Chapter 7

Om-omission

Gosse Bouma
University of Groningen

The Dutch complementizer *om* is optional if the clause it introduces is a complement. We show that a large part of the variation in the distribution of *om* is accounted for by the governing verb. Syntactic complexity also plays a significant role, as well as semantic properties of the embedded clause.

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, but here we will consider verbs only:

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

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)) has a very strong effect on the probability that the IC is introduced by *om*. Another factor that might play a role is processing complexity. Processing complexity 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

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:

(2) 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 (IJbema 2002). 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.

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.[1] However, there has not been any corpus-based study into the distribution of *om* that investigates the features that influence the presence or absence of *om* in individual sentences. This is in strong contrast with a similar phenomenon in English, i.e. the optional presence of *that* in finite complement clauses, which has been the subject of numerous studies (see, among others, Ferreira & Dell (2000) and Hawkins (2002)). In particular, Roland, Elman & 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 general.

2 Why add *om*?

There are two considerations that might explain why language users sometimes do and sometimes don't include *om*: processing complexity and semantics. A complementizer explicitly marks the beginning of an infinitival clause, and as such can help to reduce processing complexity. Roland, Elman & Ferreira (2006) observe that in English 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. Jaeger (2010) gives a similar but more general account in terms of *information density*. One might argue that choice for the complementizer *om* in Dutch can be explained in a similar way. Furthermore, 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 as well.

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

[1] A comparison between the Corpus of Spoken Dutch and the newspaper corpus used in this study confirms that *om* is indeed more frequent in spoken language.

om as complement express a meaning that makes the complement clause very close in meaning to a purpose or goal clause:

(3) De EU zal alles in het werk stellen om te helpen
The EU will everything in the work put COMP to help
'The EU will do everything it can to help.'

A semantic account predicts that complement clauses that are close in meaning to a goal or purpose clause, will more likely be introduced by *om*. The opposite idea is to measure how typical a combination of matrix verb and (the head of) an IC is. ICs headed by verbs that are 'typical' for a given matrix verb are probably less likely to be introduced by *om*.

3 Data

We used an 80 million word subset of the Twente Newspaper corpus (Ordelman et al. 2007) as corpus.[2] For computing semantic association scores, we used the full Twente Newspaper corpus (500 million words). The corpus was parsed automatically using Alpino (van Noord 2006). 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 *om*. We took several measures to ensure that the amount of noise is kept to a minimum.

Initially, we selected all sentences containing a TI (*te-infinitival*) or OTI (*om-te-infinitival*) clause functioning as verbal complement, i.e. with grammatical relation label VC in the dependency graph output by the parser.[3] We filter all examples involving governors that did not 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*):

(4) ...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.'

(5) ...waarna hij **besloot** (om) zich blijvend in de VS te vestigen

Example (4) exhibits cross-serial dependency word order where insertion of *om* is never possible. In (5), 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.

[2] Consisting of material from *Algemeen Dagblad* and *NRC Handelsblad*, 1994 and 1995.
[3] Subject TI and OTI clauses are rare and were ignored.

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*).

4 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.

23% of the 49K relevant examples in our corpus contains *om*. There are 95 different verbal governors, 23 of which have a preference for OTI over TI. 11 verbal governors occur with an OTI less than 10% of the time. 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 1 (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. On the other hand, the right pane of Figure 1 shows that the correlation between the frequency with which a verb occurs with a verbal complement and the probability of complementizer presence (as observed for English) does not hold for our data.

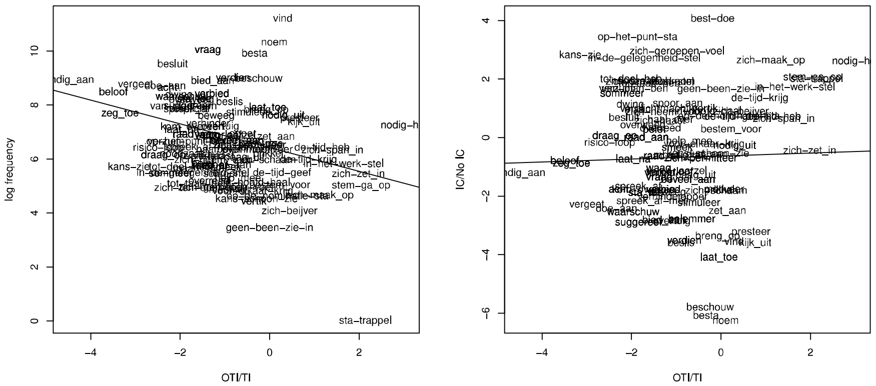

Figure 1: Overall log frequency of a verbal governor against ratio of OTI occurrence (left pane) and ratio of overall IC over non IC occurrence of verbal governors against ratio of OTI occurrence (right pane).

We 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 syntactic complexity: length of the IC (in number of words), relative position of the *te*-infinitive heading the IC from the start of the IC, syntactic category of the first constituent of the IC (*nominal, adverbial, verbal*, or *other*).

Table 1 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.

Table 1: Percentage OTI

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	1694	1108	40.0

(a) %OTI for various distances between between governor and start of the 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.

We can also look at the category of the clause headed by the verbal governor. If this is a finite main clause, we expect the percentage of *om* to be higher. Table 1b 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.

Complexity can also be caused by the presence of other complements in the matrix clause. Our data shows that the probability of OTI goes up strongly if an inherent reflexive (45.6% OTI), predicative complement (83.8%), or expletive *het* (50.3%) 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. Using binary features that measure the presence of such complements can be an alternative for using valence frames.

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. 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 & Hanks 1990) between the two (where $f(\text{W})$ is the relative frequency of W in the corpus:

$$\text{pmi(Gov,IC-head)} = \ln\left(\frac{f(\text{Governor,IC-head})}{f(\text{Governor})\cdot f(\text{IC-head})}\right) \tag{6}$$

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:

$$\text{pmi(PurposeClause,Head)} = \ln\left(\frac{f(\text{PurposeClause,Head})}{f(\text{PurposeClause})\cdot f(\text{Head})}\right) \tag{7}$$

To obtain the relevant statistics, we assume that all OTI constituents in the corpus that have the dependency relation MOD express a purpose or goal. 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'.

5 Experiments

We describe experiments to determine which properties influence the choice for *om*, and how these properties interact. We used R and *lme4* (Bates, Maechler & Bolker 2011) to perform a linear mixed effects analysis, where verbs are random effects (see Baayen 2008).

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. 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 (Table 2, *Main clauses only*) includes distance between governor and IC (*dist*), length of the TI, distance between start of the TI and the *te*-infinitive verb (*te*), and presence of expletive *het* (*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

Table 2: Best model using verbal sense of the governor as random effect and various syntactic complexity features as fixed effects.

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

	Main clauses only			All clause types		
	effect	std. err	significance	effect	std. err	significance
(Intercept)	-0.90	0.20	***	-0.98	0.13	***
dist	0.13	0.05	*	0.15	0.02	***
length	-0.13	0.03	***	-0.10	0.02	***
te	0.27	0.04	***	0.20	0.02	***
het	0.38	0.19	*	0.49	0.12	***

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[4] = 15,716, Baseline AIC = 16,001, χ^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.

Next, 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 2 (*All clause types*). The model outperforms the baseline significantly (χ^2=549.88, p < 0.001, Model AIC = 40,087, baseline AIC = 40,601). 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.

To test our hypothesis that semantics might play a role, we use two features based on pointwise mutual information, as explained in Section 4. A model that uses only these two features as fixed effect is given in Table 3. 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 (χ^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. A model using both complexity

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

Table 3: Model and fixed effects for the complete dataset using semantic features.

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

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

features and semantic features does perform better than the model using complexity features only (χ^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 predictive qualities.[5] We conclude that complexity and semantic factors both influence the choice for *om*.

6 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. The matrix verb influences the likelihood of *om* significantly and thus we decided to use a mixed effects model, where the verb is used as random effect. Features that reflect syntactic complexity play a significant role. 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 as well, 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: manually corrected treebanks might give rise to more accurate data and stronger effects, medium and genre is likely to play a role,[6] but requires a balanced corpus, and finally, other measures for syntactic complexity (such as local and global sentence ambiguity according to a parser) could be explored.

References

Baayen, R. H. 2008. *Analyzing linguistic data.* Cambridge University Press.

[5] 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.

[6] Indeed, we found that including the source newspaper as factor already had an effect.

Bates, Douglas, Martin Maechler & Ben Bolker. 2011. *lme4: linear mixed-effects models using S4 classes*. R package version 0.999375-39. http://CRAN.R-project.org/package=lme4.

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

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

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*. Leiden University PhD thesis.

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.

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

Roland, D., J. L. Elman & 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 & Patrick Watrin (eds.), *TALN06. Verbum ex machina. Actes de la 13^{e} Conference sur le Traitement Automatique des Langues Naturelles*, 20–42.