Nederlandse Taalkunde
www.nederlandsetaalkunde.nl
Uitgave: Amsterdam University Press

Control, causation and Google counts[1]

Natalia Levshina

NT 21 (2): 253–263
DOI: 10.5117/NEDTAA2016.2.LEVS

Abstract
This contribution is a review of Chapter 5.2 of Broekhuis & Corver's (2015) *Syntax of Dutch* (pp. 765-935). The chapter describes Dutch infinitival constructions from a formal perspective. In this review I share several theoretical, descriptive and methodological considerations, which relate to the status of control as a purely syntactic phenomenon, the semantic and geographic variation of Dutch causative constructions, and the problems with using Google counts from the World Wide Web when writing a (formal) grammar.

Keywords: infinitival constructions, control, causatives, World Wide Web

1 Introduction

The chapter that I reviewed (Broekhuis & Corver 2015: 765-935) deals with different types of Dutch infinitival constructions that serve as arguments to other verbs. The main types are constructions with optional complementizer *om* and infinitival marker *te*, as in (1a); constructions with the *te*-infinitive, as in (1b), and constructions with bare infinitives, as in (1c).

(1) a. Jan beloofde Marie (om) dat boek te lezen. (p. 776)
Jan promised Marie COMP that book to read.
'Jan promised Marie to read that book.'

1 I wish to thank Professor Barbara Stiebels from Leipzig University for an enlightening discussion of control phenomena. Of course, all misconceptions are solely mine.

b. Marie beweerde dat boek te lezen. (p. 803)
Marie claimed that book to read.
'Marie claimed to be reading that book.'
c. Jan zag de kinderen lachen. (p. 905)
Jan saw the children laugh.
'Jan saw the children laugh.'

Describing these constructions, the authors discuss a range of very different phenomena: from control and subject raising to deontic and epistemic modality, and from argument clause transparency to direct and indirect causation. Therefore, it is quite a challenge to do justice to every topic discussed in the chapter. In what follows I will focus on three aspects: the problem of syntactic control, variation in Dutch analytic causatives and the adequacy of Google frequency information. I will also present a small pilot study that deals with the use and omission of the optional complementizer *om* because, in my opinion, this aspect is not discussed in the chapter in sufficient detail.

2 Control: how much syntax do you need?

In this section I will focus on the variation of *te*-infinitivals with and without the optional complementizer *om*. These two cases are illustrated by the examples in (1a) and (1b), respectively. With a very high degree of simplification, one can summarize the main ideas of the corresponding subsections as shown in Table 1 (cf. pp. 814-815), which also displays several examples of the matrix verbs that combine with the infinitivals.

Table 1 Syntactic and semantic features of two types of infinitival clauses

	Syntax	Semantics	Examples of matrix verbs
(om) + *te*-infinitivals	non-obligatory control	non-propositional	*afkeuren* 'disapprove of', *beloven* 'promise', *besluiten* 'decide', *verzoeken* 'request', *weigeren* 'refuse'
te-infinitivals	obligatory control	propositional	*beseffen* 'realize', *beweren* 'claim', *verdenken* 'suspect', *verweten* 'reproach', *verzekeren* 'assure'

From a syntactic perspective, (*om*) + *te*-infinitivals are non-obligatorily controlled, whereas *te*-infinitivals are either obligatorily controlled, as in (2), or involve subject raising, as in (3).

(2) Jan beweert de beste te zijn. (p. 844)
Jan claims the best to be.
'Jan claims to be the best.'

(3) Jan blijkt dat boek te lezen. (p. 847)
Jan turns.out that book to read.
'Jan turns out to be reading that book.'

Generally speaking, complement control is observed when a verb takes a state of affairs argument (e.g. an infinitival clause) and an argument of the matrix predicate is identified with an argument of the embedded predicate (Stiebels 2007). For example, in (4a) the subject of the matrix clause (Mary) is identified with the covert subject of the embedded clause. This covert subject is represented as PRO. In contrast, in (4b) it is the object of the matrix clause that is identified with the covert subject of the embedded clause.

(4) a. Mary_i promised John_j [$\text{PRO}_{i/*j}$ to call the doctor].
b. Mary_i asked John_j [$\text{PRO}_{*i/j}$ to call the doctor].

Obligatory and non-obligatory control differ with regard to several syntactic criteria (cf. p. 782), such as the possibility of omission of the PRO antecedent in the matrix clause, e.g. in impersonal passive constructions, the possibility of split antecedents and arbitrary interpretation of PRO. These criteria will not be discussed here due to the lack of space (but see §5.2.1.3, sub III, 782ff in the *Syntax of Dutch*). Overall, one can conclude that non-obligatory control involves a looser relationship between the covert subject in the embedded clause and its antecedent in the matrix clause and thus a weaker degree of clause integration.

From a semantic perspective, *te*-infinitival clauses are propositional, i.e. those that can be assigned a truth value, whereas (*om*) + *te*-infinitivals are non-propositional, i.e. those that cannot have a truth value and refer to potential state-of-affairs (cf. p. 815).

These observations demonstrate that the formal theory of grammar that the *Syntax of Dutch* assumes is not so different from functional theories of grammar. The chapter shows a striking correlation between form (clause integration and the possibility of *om*) and meaning (propositional or non-propositional clause). In functional linguistics and typology, such correlations have been described extensively, e.g. in Givón's (1990) binding hierarchy, Van Valin's (1993) hierarchy of predicates and Cristofaro's (2003)

deranking hierarchy. It is inspiring to watch these very different theories converge.

The authors write on p. 802 that non-obligatory control is a matter of pragmatics, rather than syntax and semantics. They conclude that after a discussion of control shift. Consider the examples in (5). The action of *komen* 'come' in (4a) is expected to be performed by Jan, whereas in (5b) the actor is Peter. In other words, (5a) is an instance of subject control, and (5b) is an example of object control.

(5) a. Jan_i beloofde Peter_j (om) $\text{PRO}_{i/*j}$ te komen. (p. 802)
Jan promised Peter COMP to come.
'Jan promised Peter to come.'

b. Jan_i beloofde Peter_j (om) $\text{PRO}_{i*/j}$ te mogen komen. (p. 802)
Jan promised Peter COMP to be.allowed.to come.
'Jan promised Peter to be allowed to come.'

This difference has been explained pragmatically. In commissive speech acts (e.g. the one expressed by *beloven* 'promise', as in (5a)), one of the felicity conditions is that the speaker intends to perform the action that is promised. Thus, the default interpretation of sentences with commissive matrix verbs is the one with subject control. The sentence in (5b) overrides the default interpretation because it is unusual to grant permission to oneself. Importantly, control shift is related to such fundamental phenomena as markedness and the Principle of Economy: the expression in (5b) contains more overt coding material than the one in (5a) because it conveys a more 'marked', i.e. less expected meaning (Comrie 1986).

Notably, some *te*-infinitivals, which are considered to involve obligatory control, exhibit similar variation, e.g. *verwijten* 'reproach', as in (6):

(6) a. Jan_i verweet haar_j [PRO_j niets te doen]. (p. 810)
Jan reproached her nothing to do.
'Jan reproached her for not doing anything.'

b. Jan_i verweet haar_j [PRO_i niets te mogen doen]. (p. 810)
Jan reproached her nothing to be.allowed do.
'Jan reproached her for not being allowed to do anything.'

The mechanism is similar to the one discussed in the example with *beloven* 'promise'. Following the line of reasoning on p. 802, does this mean that obligatory control is also a matter of pragmatics (cf. Duffley 2014)? I understand (and appreciate) the authors' reluctance to go into purely theoretical

discussions, but what is the borderline between pragmatics and syntax then, in their view, and does one actually need it in order to produce an adequate linguistic description?

Remarkably, from a typological perspective, control shift seems to be mainly a (West)-Germanic phenomenon. Previous research shows that German allows for control shift more often than English (see Stiebels 2007 for an overview). Considering the vast existing research on the so-called *Germanic sandwich*, which puts Dutch between English and German with regard to numerous linguistic phenomena (e.g. van Haeringen 1956), one might wonder whether the syntactic and semantic conditions and usage frequencies of control shift in Dutch are intermediate between English and German.

Another question that I had in my mind while reading the chapter was the following: What explains variation in the use of the optional complementizer *om*? The authors discuss the variation very briefly. They say, namely, that the omission of *om* may make the embedded clause transparent (p. 776). In other words, its elements can be used before the matrix verb, as *dat book* 'that book' in (7a). If the embedded clause contains *om*, it should always follow the matrix verb in clause-final position.

(7) a. dat Jan <dat boek> weigert <dat boek> te lezen. (p. 776)
that Jan that book refuses to read.
'that Jan refuses to read that book.'

b. dat Jan <*dat boek> weigert om <dat boek> te
that Jan that book refuses COMP to
lezen. (p. 776)
read.
'that Jan refuses to read that book.'

But what triggers the use and omission of *om*? Unfortunately, this intriguing question is not addressed in the chapter. In a small and very preliminary case study based on a web-based Dutch corpus from the Leipzig Corpora Collection[2] (1 million sentences from various web sources), I found instances of several matrix verbs followed by the *te*-infinitive, with or without *om*. For highly frequent verbs, only a part of the corpus was explored, due to the fact that all examples needed to be checked manually. As shown in Table 2, *aarzelen* 'hesitate' has the highest proportion of complements with *om* (66.7%), whereas *proberen* 'try' has the lowest proportion (7.5%).

2 http://corpora.uni-leipzig.de/, accessed at 06.05.2016.

Table 2 Frequencies and proportions of *om* + *te*-infinitivals after different matrix verbs in the Dutch component of the Leipzig Corpus Collection (1M sentences)

	om + *te*	Total number of infinitivals
proberen 'try' (for 30,000 sentences only)	8 (7.5%)	107
beloven 'promise'	21 (10.4%)	202
verzoeken 'request'	103 (22.4%)	460
besluiten 'decide' (for 100,000 sentences only)	74 (30%)	246
adviseren 'recommend' (for 200,000 sentences only)	29 (37.2%)	78
verlangen (ernaar) 'long for'	6 (54.5%)	11
aarzelen 'hesitate'	62 (66.7%)	93

One may wonder what explains the differences in the percentages of *om*. A possible factor is the degree of subjectivity on the part of the matrix subject. The verbs that do not convey directly a subjective evaluation of the situation expressed in the embedded clause, such as *proberen* 'try', have lower proportions of infinitivals with *om*, whereas the verbs that foreground subjective evaluation, such as *aarzelen* 'hesitate' and *verlangen* (*ernaar*) 'long for', have higher proportions. One can thus hypothesize that *om … te …* marks a subjunctive infinitival construction. This ties in with the previous observations about the speaker-oriented meaning of *om … te …* in some contexts (Blom 1984, Vliegen 2001). Although an examination of other verbs and potential factors is needed, including the stigmatization of the redundant *om* by normativists (p. 774), these preliminary results look promising. Regardless of the theoretical status of these findings, I believe that such variation should be part of a descriptive grammar of a language.

3 Dutch analytic causatives

Dutch constructions with the bare infinitive are extremely diverse. They include constructions with modal verbs, perception verbs like *zien* 'see' and *horen* 'hear', the verbs *leren* 'learn/teach' and *hebben* 'have', and the causative auxiliaries *doen* 'do, make' and *laten* 'let'. It is the latter that I would like to focus on in this review.

The corresponding section 5.2.3.4 in the *Syntax of Dutch* provides the most important details about the Dutch analytic causatives concerning the

formal, semantic, stylistic and geographic variation of these constructions. Unfortunately, the authors do not mention a large body of corpus-based research dedicated to *doen* and *laten* (e.g. Kemmer & Verhagen 1994, Verhagen & Kemmer 1997, Stukker 2005, Speelman & Geeraerts 2009, Levshina 2011, Levshina, Speelman & Geeraerts 2013), which would help to make the description more precise, in my view. For example, the book says that Dutch speakers from Belgium are often more permissive in their acceptability judgements than speakers from the Netherlands when *doen* is used with animate Causers. However, it has been shown that Belgian Dutch speakers use *doen* more often across a variety of contexts, as shown in several multifactorial corpus-based studies where the Causer's animacy, as well as a number of other semantic and syntactic variables are controlled for (e.g. Speelman & Geeraerts 2009). Moverover, the *doen*-constructions with animate Causers acting intentionally are not typical of Standard Belgian Dutch, either (Levshina 2011).

In addition, the authors point out that *doen* is more frequently used in informal language and in fixed expressions, such as *van zich doen spreken* 'make one's mark' (lit. 'make speak about oneself'). While this observation is supported by the results of the previous corpus-based studies, it is important to mention that *laten* is also used in a number of highly frequent set expressions, such as *laten zien* 'show' (lit. 'let see'), *laten weten* 'inform' (lit. 'let know') and *een steek laten vallen* 'make a stupid mistake' (lit. 'let a stitch fall'). In particular, *laten zien* 'show', *laten weten* 'inform' and *laten horen* 'express, make (e.g. one's voice) heard' (lit. 'let hear') are highly frequent in Netherlandic Dutch (Levshina, Geeraerts & Speelman 2013). This suggests a higher level of grammaticalization of *laten* in the Netherlandic variety, where it has become the default causative marker.

An interesting feature of the Dutch causatives, especially the one with *laten*, is that the Causee can be marked with a preposition (*aan* 'to' or *door* 'by'). The chapter says that the choice between the prepositions seems to be determined by the embedded infinitive, but the exact mechanism involving individual verbs is still unclear (p. 926). However, this variation has already been discussed in cognitive linguistic literature. According to Kemmer & Verhagen (1994), the marking of the Causee depends on the transitivity of the effected predicate, the role of the Causee in the effected event and whether the causation is construed as direct or indirect. For instance, one can choose between (8a), (8b) and (8c):

(8) Kemmer & Verhagen (1994: 136)
a. Hij liet haar de brief lezen. (zero marking)
He let her the letter read.
'He let/had her read the letter.'
b. Hij liet de brief aan iedereen lezen. (DAT)
He let the letter to everybody read.
'He let/had everybody read the letter.'
c. He liet de brief door iemand lezen. (AGENT/INST)
He let the letter by somebody read.
'He had the letter read by somebody.'

The reader marked with the agentive/instrumental case in (8c) is a metaphorical instrument, who is only of incidental importance. One can imagine, for instance, that the reading is done for such purposes as checking grammar and spelling, or for some implied audience. Importantly, the reader does not read the letter for content. With the dative (8b), however, reading for content is exactly the interpretation we find. The purpose is to let the readers find out what the letter says. The difference between (8b) and (8a) is that the Causee expressed by the zero-marked nominal phrase is more topical than the one expressed by the dative preposition (Kemmer & Verhagen 1994: 136). Of course, the specific infinitives in the causative constructions are relevant for the prepositional marking, but only to the extent that they can constrain the role repertoire of the Causee. This alternative view deserves at least to be mentioned, in my opinion.

4 Do Google counts count?

Although formal linguists traditionally rely on introspection, the authors of *Syntax of Dutch* quite often provide counts from the World Wide Web returned by the Google search engine. It has been argued that Google is a valid source of linguistic data and that its results are comparable to ones obtained from standard corpora (cf. Taylor 2012: 17-18). However, there are a few caveats, which are described below.

First, the book aims at describing Standard Dutch, as we learn from the Introduction. At the same time, we cannot know the linguistic profile of the users and whether they are using dialect, regiolect, *tussentaal* or another variety. We also do not know how the sentences were produced (i.e. whether machine translation software was used). Since there exist several very large, morphologically and syntactically annotated corpora of Stan-

dard Netherlandic and Belgian Dutch (e.g. Twente Nieuws Corpus and Leuven Nieuws Corpus), the authors' choice is not obvious.

Second, using Google is problematic if one wants to obtain exact absolute counts because some documents can be reported more than once. For low-frequency constructions – and those are the ones in which generativist syntacticians are usually interested – this may cause a distortion.

A third problem is the reproducibility of the results. For instance, on p. 786 the authors report the results of a Google search (15.11.2012) on the strings [*er werd/is geprobeerd om*] 'it was/has been tried' and [*er werd/is vergeten om*] 'It was/has been forgotten'. They write that the search resulted in more than one million hits for both cases. I have tried to reproduce these results. My search (30.04.2016) yields the following counts:

– *er werd geprobeerd om*: about 8,000 results;
– *er is geprobeerd om*: about 16,700 results;
– *er werd vergeten om*: 383 results;
– *er is vergeten om*: about 4,350 results.

Since it is very unlikely that the usage of Dutch impersonal passives has changed so radically over four years, one can suspect that there are differences in the method. My guess is that the authors did not perform searches for exact phrases, as I did. If this is the case, many examples that the authors found must have been spurious. Consider an example in (9), which was returned as a result of the search for *er werd geprobeerd om* (not an exact expression):

(9) We hebben **geprobeerd om** iedereen tevreden te stellen en aan zijn of haar … **Er werd** gestreden **om** elk punt en de meeste wedstrijden gingen ook door tot het … (http://www.dvodarts.nl/nieuwsartikelen?start=20, retrieved on 05.05.2016)

However, even if all counts are perfectly correct or even retrieved from a representative standard corpus, there is a deeper issue, I believe. In contrast to usage-based approaches to language, the generative theories of grammar presuppose that the grammatical knowledge is non-probabilistic. The question is then, how many Google examples should be sufficient to claim that a construction is part of the speaker's language? Consider a passage on p. 814: 'And when we check the internet for the string [object pronoun + *gezegd om te*], we indeed find a sufficiently large number of examples with the intended directive meaning to warrant the claim that

we are in fact dealing with *om* + *te*-infinitivals.' The counts are not provided, unfortunately, but even if they were shown, it is not clear what the sufficient number of examples is in order to support a theoretical claim. In other words, how many counts count?

5 Conclusions

In general, the chapter provides many interesting and important observations that can be useful to advanced students and professional linguists of different theoretical persuasions. The book is written clearly, without much professional jargon. I find particularly helpful the presence of separate subsections with conclusions and discussion overviews. The book triggers many new questions and whets appetite for finding out more about the syntax of Dutch.

Despite all these obvious merits, I find the lack of references to relevant works outside formal grammar deplorable, especially in the cases when semantic, geographic and other variation is discussed. Secondly, the use of Google counts is problematic in many methodological and theoretical ways. Finally, considering the vast body of typological literature on complementation, a desideratum for future reference works on Dutch infinitival complements would be to add a systematic cross-linguistic perspective, showing how the use and structure of the Dutch constructions reflect the universals of language evolution.

References

Blom, Alied (1984). Enkele opmerkingen over *te* en *om*. *Spektator* 14, 170-176.

Broekhuis, Hans & Norbert Corver (2015). *Syntax of Dutch: Verbs and Verb Phrases Volume II*. Amsterdam: Amsterdam University Press.

Comrie, Bernard (1986). Markedness, grammar, people, and the world. In: Fred R. Eckman, Edith A. Moravcsik & Jessica R. Wirth (eds.), *Markedness*. New York: Plenum Press, 85-106.

Cristofaro, Sonia (2003). *Subordination*. Oxford: Oxford University Press.

Duffley, Patrick (2014). *Reclaiming Control as a Semantic and Pragmatic Phenomenon*. Amsterdam: John Benjamins.

Givón, Talmy (1990). *Syntax: a functional-typological introduction*. Vol. II. Amsterdam: John Benjamins.

Haeringen, Coenraad B. van (1956). *Nederlands tussen Duits en Engels* [Dutch between German and English]. Den Haag: Servire.

Kemmer, Suzanne & Arie Verhagen (1994). The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* 5, 115-156.

Levshina, Natalia (2011). *Doe wat je niet laten kan* [*Do what you cannot let*]*: A usage-based analysis of Dutch causative constructions*. PhD dissertation, University of Leuven.

Levshina, Natalia, Dirk Geeraerts & Dirk Speelman (2013). Towards a 3D-Grammar: Interaction of linguistic and extralinguistic factors in the use of Dutch causative constructions. *Journal of Pragmatics* 52, 34-48.

Stiebels, Barbara (2007). Towards a typology of complement control. In: Barbara Stiebels (ed.), *ZAS Papers in Linguistics 47. Studies in complement control*, 1-80.

Speelman, Dirk & Dirk Geeraerts (2009). Causes for causatives: the case of Dutch 'doen' and 'laten'. In: Ted Sanders & Eve Sweetser (eds.), *Causal categories in discourse and cognition*. Berlin/New York: Mouton de Gruyter, 173-204.

Stukker, Ninke (2005). *Causality marking across levels of language structure*. PhD thesis, University of Utrecht.

Taylor, John (2012). *The Mental Corpus: How Language is Represented in the Mind.* Oxford: Oxford University Press.

van Valin, Robert D. Jr. (1993). A synopsis of Role and Reference Grammar. In: Robert D. van Valin Jr. (ed.), *Advances in Role and Reference Grammar*. Amsterdam: John Benjamins, 1-164.

Verhagen, Arie & Suzanne Kemmer (1997). Interaction and Causation: Causative Constructions in Modern Standard Dutch. *Journal of Pragmatics* 27, 61-82.

Vliegen, M.L.M.J. (2001). Het facultatieve 'om' na illocutionaire werkwoorden. *Nederlandse Taalkunde* 2, 112-132.

About the author

Natalia Levshina, Leipzig University
E-mail: natalia.levshina@uni-leipzig.de