A treebank-driven investigation of predicative complements in Dutch

An efficient, practical, actually usable approach

Frank Van Eynde

Centrum voor Computerlinguïstiek, Universiteit Leuven

Abstract

Treebanks are used for various purposes in language technology, but the wealth of data they contain can also be put to good use for the purpose of linguistic description and linguistic theory. To demonstrate this I will show how the treebank of the Spoken Dutch Corpus can be exploited to improve our understanding of what it is that distinguishes predicative complements from other types of complements. Section 1 shows why this distinction matters, section 2 provides a brief presentation of the treebank, section 3 gives a comprehensive survey of the intransitive predicate selecting verbs, based on the treebank data, section 4 presents a number of factors which can be used to differentiate the predicate selecting uses of the relevant verbs from their other uses, and section 5 summarizes the results.[1]

1 Predicative complements

Distinguishing a predicative complement from an object complement is easy in pairs like (1).

(1) a. Fred is a plumber.
b. Fred knows a plumber.

The complement of the copula denotes a property which is attributed to the referent of the subject, and the copula itself is little more than a carrier of mood and tense. By contrast, the complement of *know* denotes an entity and the verb denotes a binary relation between that entity and the referent of the subject.

Since the verb is the only element that overtly distinguishes (1a) from (1b), it might seem sufficient to draw the distinction, but the matter is more complex, since many verbs are used either way. The second complement of *call* and *make*, for instance, is predicative in (2), but not in (3).

(2) a. Don't call me a liar.
b. They will make you chairman.

[1]This work is part of a larger project on the syntax and semantics of clauses with predicative complements. So far, it has yielded an HPSG style analysis of such clauses (Van Eynde 2008) and a semantic analysis of the copula, presented at the HPSG-2009 conference.

Proceedings of the 19th Meeting of Computational Linguistics in the Netherlands
Edited by: Barbara Plank, Erik Tjong Kim Sang and Tim Van de Cruys.

(3) a. Please call me an ambulance.
b. They will make you a cake.

This shows that the relevant distinction is that between the predicate selecting uses of the verbs and their other uses. The aim of this paper is to identify and classify the predicate selecting verbs (section 3) and to provide criteria for distinguishing their predicate selecting uses from their other uses (section 4). For this purpose we can get a lot of mileage from a treebank. The one I will employ is that of the Spoken Dutch Corpus (Corpus Gesproken Nederlands) (section 2).[2]

2 The CGN treebank

The Spoken Dutch Corpus (CGN) contains approximately 1000 hours of speech, which roughly corresponds to 10 million words (Oostdijk 2000). Two thirds were recorded in the Netherlands and one third in Flanders, the Dutch speaking part of Belgium. All of the recordings have been transcribed and syntactically annotated, following the guidelines of the annotation manual (Hoekstra et al. 2003). The annotation takes the form of directed acyclic graphs with information about constituency and dependency, as exemplified in Figure 1. Every word is assigned a lexical category, such as BW for the adverbs *dan* and *wel*,[3] every phrase is assigned a phrasal category, such as SMAIN for main clauses, and the edges have a dependency label, such as HD for heads and MOD for modifiers. In the trees, the lexical categories are given just below the word, the phrasal categories are in ovals and the dependency labels in rectangles.

Part of the CGN treebank, roughly 10 %, was manually verified and corrected, where needed. Since this part is obviously more useful for the present purpose than the unverified and uncorrected data, I will only use the former. Moreover, since I used the first release of the treebank, in which only the Flemish data had been verified, the sample exclusively contains the Flemish data.[4] It consists of 42750 sentences.[5] Their representations jointly contain 382101 tokens, 216485 phrasal categories and 453312 dependency labels. Table 1 gives the frequency in the sample of a subset of the dependency labels.

Of special interest in this context are the constituents with the PREDC label. They include the complements of the copular verbs, as listed in Haeseryn et al. (1997, 1122-4):

[2]Other studies which have exploited the CGN treebank for the purpose of linguistic description include Van der Wouden et al. (2003) and Bouma (2004).

[3]The labels for the lexical categories are abbreviations of Dutch terms; BW, for instance, stands for *bijwoord* 'adverb' and WW for *werkwoord* 'verb'.

[4]In the second release, also the data from the Netherlands have been verified and corrected.

[5]For comparison, Sarkar and Zeman (2000) employs a sample of 19126 sentences from the Prague Dependency Treebank for extracting subcategorization frames for Czech, Kupsc and Abeillé (2008) employs a sample of about 20000 sentences from the Paris7 Treebank for extracting subcategorization frames for French, and Hinrichs and Telljohan (2009) employs a sample of approximately 36000 sentences from the Tüba-D/Z treebank for extracting subcategorization frames for German. A sample of 42750 sentences is, hence, comparatively large for this type of investigation.

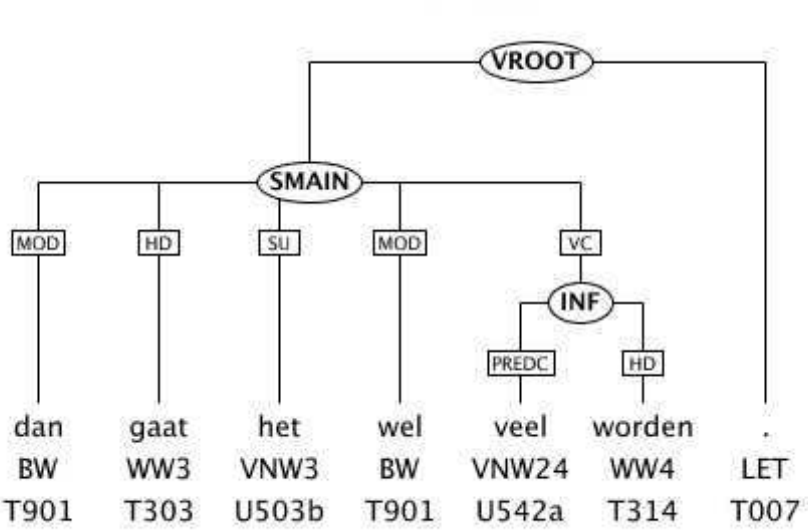

Figure 1: 'It is going to become a lot then.'

HD	head	121078
MOD	modifier	62381
OBJ1	first or direct object	39398
SU	subject	36664
VC	verbal complement	14170
PREDC	predicative complement	9173
LD	location or direction complement	4563
PC	prepositional object	4109
PREDM	predicative modifier	954
OBJ2	secondary or indirect object	736

Table 1: A subset of the dependency labels with their frequency in the sample

- *zijn* 'be', *worden* 'become' and their equivalents;
- *blijven* 'stay, remain';
- *blijken, dunken, heten, lijken, schijnen, voorkomen* 'seem, appear, be reputed, be called'.

The equivalents of *zijn* include *vallen* 'fall', *zitten* 'sit' and *staan* 'stand', as used in (4a); those of *worden* include *gaan* 'go', *komen* 'come', *lopen* 'run' and *(ge)raken* 'get', as used in (4b).

(4) a. De rivier staat al meer dan een maand droog.
the river stands already more than a month dry
'The river has been dry for more than a month now.'

b. dat hij soms in de war (ge)raakt.
that he sometimes in the confusion gets
'that he is sometimes getting confused.'

The PREDC label is also assigned to the constituents which are known as *bepaling van gesteldheid*. This term was coined by 19th century Dutch grammarians and is left untranslated here, since it has no equivalent in the grammars of other languages. As it subsumes a rather heterogeneous class of constituents, it was partitioned by 20th century grammarians into a number of subtypes, the basic distinction being that between adjuncts and complements (van den Toorn 1969). An example of the former is the adjective in *Ze vingen de leeuw levend* 'they caught the lion alive'. In the treebank they are assigned the dependency label PREDM and for the purpose of this investigation they can be ignored, since we only deal with predicative complements. An example of the latter is the adjectival complement in (5).

(5) Dat voorstel klinkt heel interessant.
that proposal sounds very interesting
'That proposal sounds very interesting.'

It denotes a property which is attributed to the referent of the subject, just like the complements of copular verbs. In combination with a transitive verb the predicative complement can also denote a property which is attributed to the referent of the direct object, as in (6).

(6) Ik vind de soep heerlijk.
I find the soup delicious
'I consider the soup delicious.'

In contrast to what is usually done for the copular verbs, Dutch grammars do not provide much detail about the class of verbs which select a *bepaling van gesteldheid*. It is left to the reader to extrapolate from a few clear-cut cases to all relevant cases.

3 The selecting verb

For the purpose of natural language processing the current treatment in descriptive grammars, as sketched above, does not provide a solid starting point. With its open-ended list of 'equivalents' of the copular verbs and its simple enumeration of examples of *bepaling van gesteldheid* it is not sufficiently precise and detailed for inclusion in a parser, for instance. As a first step in the direction of a more precise and comprehensive treatment I will identify and classify the set of verbs that take a predicative complement in Dutch, employing the wealth of data that is included in the CGN treebank.

Given the limitations on the size of contributions for these proceedings, the section does not cover all of the predicate selecting verbs, but only the **intransitive** ones. Characteristic of these verbs is that their predicative complement denotes a property which is attributed to the referent of the subject, also if they have an object complement, such as the experiencer denoting pronouns in (7).

(7) a. Arthur lijkt ons een geschikte kandidaat.
arthur seems us an appropriate candidate
'Arthur seems an appropriate candidate to us.'

b. De situatie wordt me hier te lastig.
the situation becomes me here too tricky
'The situation is getting too tricky for me here.'

Another common property is that they cannot be passivized, also if the predicative complement is a nominal, as in (8).

(8) * Journalist wordt niet door iedereen geworden.
* journalist is not by everyone become

There are, of course, many ways to partition this class of verbs. The one I found most useful as a starting point is that between stative and dynamic verbs. Representative members are respectively *blijven* 'remain, stay' and *worden* 'become'. In model-theoretic terms, they are duals. To become dry, for instance, is not to stay not dry, i.e. not to stay wet, and to stay dry is not to become not dry, i.e. not to become wet. In other words, *worden* is the denial of permanence, and *blijven* is the denial of change. A syntactic test is the admissibility in the infinitival *aan het* construction. Since the *aan het* construction presents a situation as evolving in time, it is compatible with the dynamic *worden* 'become', but not with the stative *blijven* 'stay'.

(9) a. Ze was in ijltempo volwassen aan het worden.
she was in fast-rate adult at the become
'She was becoming an adult in no time.'

b. * Ze zijn volwassen aan het blijven.
* they are adult at the stay

Both types will be discussed and further partitioned in the following paragraphs.

3.1 The stative intransitive predicate selectors

Table 2 provides a survey of the stative predicate selectors in the CGN sample. For each of them the PREDC column specifies how often they are combined with a predicative complement.[6] Notice that the numbers concern the **intransitive** uses. The transitive predicate selecting uses which some of the listed verbs have, as in (10), are not included.

(10) a. Ze wil hem weg.
she wants him away
'she wants him away'

b. Ik zie dat als een bedreiging.
I see that as a threat
'I see that as a threat.'

The table also specifies for each verb how often it occurs in the sample, see the columns called 'Total'. The fact that these numbers are systematically higher, and –in some cases– much higher than those in the PREDC columns demonstrates that all of the predicate selectors are also used in other ways. One of those ways concerns the combination with a complement of location or direction (LD). The numbers for this combination are specified as well, since they provide the basis for a finer-grained classification, differentiating the predicate selectors that also combine with such complements from those which do not. The former are in the left half of the table and the latter in the right half.

Orthogonal to this dichotomy is another one which concerns the combination with infinitival complements. More specifically, the predicate selecting uses of the verbs in the lower half of the table result from the omission of an infinitival complement.

(11) a. Het kan niet beter (zijn).
it can not better (be)
'It couldn't be better.'

b. dat die oplossing ons geschikt lijkt (te zijn).
that that solution us appropriate seems (to be)
'that that solution seems appropriate to us.'

If the infinitive is there, the adjectival predicate is a complement of the infinitive, but if it is absent, the adjective is a complement of the matrix verb. Characteristic of these verbs is that they lack the imperative.

(12) a. * Moet/mag/kun beter!
* must/may/can better!

[6] *Dunken* is not mentioned, since it does not have a single predicate selecting use in the sample. The same holds for the stative *vallen*, as used in *Het afscheid valt ons zwaar* 'the goodbye is emotionally heavy for us'.

	PREDC	LD	Total			PREDC	Total	
zijn	7193	800	11919	be	*er uit zien*	27	31	look
blijven	74	38	296	remain	*zien*	2	926	look
staan	77	445	667	stand	*klinken*	23	31	sound
zitten	48	450	752	sit	*aanvoelen*	1	5	feel
liggen	37	121	246	lie	*smaken*	2	4	taste
hangen	7	40	100	hang	*ruiken*	1	7	smell
					overkomen	4	15	come across
					voorkomen	1	39	appear
kunnen	15	48	1941	can	*lijken*	40	89	seem
moeten	4	47	1978	must	*blijken*	4	39	appear
mogen	1	17	430	may	*schijnen*	3	42	seem
willen	5	14	631	want	*heten*	21	61	be called
hoeven	1	0	31	need				
Sum	7462	2020				129		

Table 2: The stative intransitive predicate selectors in the sample

b. * Schijn/lijk/blijk toch wat meer geinteresseerd!
   * seem/appear – what more interested!

By contrast, the predicate selecting uses of the verbs in the upper half of the table do not result from the omission of an infinitival complement, and most of them can be used in the imperative.

(13) Wees/blijf kalm!
be/stay calm!
'Be/stay calm!'

The combination of the two dichotomies yields four classes. The first comprises the copula, *blijven* and the position verbs. They all combine with locative complements, and the copula also combines with directional complements.[7]

(14) a. Ze zijn/zitten/blijven thuis.
they are/sit/stay home.
'They are/stay at home.'

b. Ze zijn naar huis.
they are to house
'They're going home.'

The second class is that of the deontic modals. They also combine with directional complements, but not with locative ones.

[7] In this respect the Dutch copula differs from the English one.

(15) a. Ze moeten/willen/mogen naar huis.
they must/want/may to house
'They must/want/may go home.'
b. * Ze moeten/willen/mogen thuis.
* they must/want/may home

The third class is that of the sensory verbs. There are two for visual impressions (*zien, er uitzien*), one for each of the other senses, and the holistic *overkomen* 'come across'.[8]

(16) a. Anja zag bruin. (sfv400350-23)
Anja saw brown
'Anja had a tan.'
b. Bij jou voelt dat altijd zo stroef aan. (sfv400269-71)
with you feels that always so awkward
'With you it always feels so awkward.'

This class also includes the predicate selecting *voorkomen*, as used in (17).

(17) Die man komt me bekend voor.
that man comes me familiar for
'That man looks familiar to me.'

The fourth class is that of the evidential copulars. That they cannot take an LD complement is demonstrated by (18).

(18) a. Ze bleken in Amsterdam *(te zijn).
they proved in Amsterdam *(to be)
'They proved to be in Amsterdam.'
b. Ze schijnen naar Brussel *(te zijn).
they seem to Brussels *(to be)
'They seem to be going to Brussels.'

The fact that these sentences are only well-formed if the infinitive is present shows that the PPs can be a complement of *zijn*, but not of *bleken* or *schijnen*.[9]

3.2 The dynamic intransitive predicate selectors

The dynamic predicate selectors come in two types: telic and atelic. Representative members are respectively *worden* and *doen*. *Gek worden* 'become mad', for instance, denotes a transition from not being mad to being mad, while *gek doen* 'do mad' denotes an ongoing activity: the madness is presented as a property which is manifest all along. The difference correlates with the choice of the perfect auxiliary: *zijn* for the telic *worden* vs. *hebben* for the atelic *doen*.

[8]The sfv numbers refer to those in the treebank. *Ogen*, as used in *Zijn palmares oogt indrukwekkend* 'his cv looks impressive', also belongs to this class, but it does not occur in the sample.
[9]The same holds for the English equivalents, as illustrated by the contrast between the well-formed *Lee seems out of his mind* and the ill-formed *Lee seems out of town* (Pollard and Sag 1994, 104).

	PREDC	LD	Total			PREDC	Total	
(ge)raken	23	21	68	get	*worden*	277	1157	become
gaan	100	378	2413	go				
komen	45	353	1025	come				
lopen	14	74	165	run				
vallen	12	29	156	fall				
Sum	194	855				277		

Table 3: Telic intransitive predicate selectors

	PREDC	Total	
doen	26	1350	do
zich gedragen	2	3	behave
zich voordoen	1	10	pretend
Sum	29		

Table 4: Atelic intransitive predicate selectors

Besides *worden*, the telic predicate selectors include *(ge)raken* 'get' and the motion verbs *gaan* 'go', *komen* 'come', *lopen* 'run', and *vallen* 'fall'.[10]

(19) Hij viel plots in slaap.
he fell suddenly in sleep
'He suddenly fell asleep.'

The finer-grained partition which is based on the compatibility with LD complements differentiates *worden* from the other verbs. See Table 3.

The atelic predicate selectors include *doen* 'do' and the inherently reflexive *zich voordoen* 'pretend' and *zich gedragen* 'behave oneself'. The adjectival predicate in (20), for instance, expresses 'a way of being'.

(20) dat ze de laatste tijd zo geheimzinnig doet.
that she the last time so secretive does
'that she is being so secretive lately.'

They are not compatible with an LD complement. See Table 4.[11]

3.3 Summing up

The stative, telic and atelic predicate selectors in the respective tables jointly account for 8091 of the predicate selecting uses. Given that 9073 of the 9173 PREDCs

[10] In contrast to its stative counterpart, the dynamic *vallen* does occur in the sample.

[11] *Doen* is also used as a transitive predicate selector, as in *hij doet de ramen dicht* 'he closes the windows', but those uses are not included in the table.

in the sample have a verbal head sister, this amounts to 89.18 % of the predicate selecting verbal uses.[12] Most of the rest concerns uses of transitive predicate selectors.

The resulting classification of predicate selecting verbs is considerably more comprehensive and more detailed than what is standardly offered in descriptive grammars. This is at least partly due to the availability of the CGN treebank. It has enabled us to identify the relevant verbs and it has provided useful data for their classification, such as the compatibility with complements of location and direction.

4 Differentiating the predicate selecting uses from the other uses

While the lemma of the selecting verb is an important factor for differentiating the predicative complements from the object complements, it cannot be the only factor, since the verbs with predicate selecting uses are also used in other ways. A comprensive survey of the relevant disambiguating factors is beyond the scope of this paper. Instead, I will focus on two which are readily amenable to a treebank-driven investigation, i.e. subcategorization and morpho-syntactic selection.

4.1 Subcategorization

Verbs which select a complement routinely require it to belong to some specific syntactic category, when the complement is an object. The transitive *kennen* 'know' and *bezitten* 'own', for instance, require their object to be nominal, rather than adjectival, prepositional or verbal. By contrast, when the selected complement is predicative, the verb is considerably less demanding. The copula, for instance, subcategorizes for an XP where X stands for any of N, A, P or V.

Since the treebank makes the relevant distinctions for both the lexical and the phrasal categories, this claim can be put to the test. Table 5 shows the result. It specifies for each of the intransitive predicate selectors how often they combine with a complement of some given syntactic category.[13] The numbers in the column 'Sum' are identical to the numbers in the PREDC columns of the previous tables.

The verbs in the upper part of Table 5 are compatible with a nominal predicate, whereas those in the lower part are not. Interestingly, the former are to a large

[12] The remaining 100 PREDCs have either no head sister or a non-verbal one, such as the absolutive *met* 'with'.

[13] In terms of the CGN labels N comprises the lexical categories for common nouns (N[1-4]), proper nouns (N[5-8]), pronouns (VNW[1-27]) and cardinals (TW1), as well as the phrasal category NP. A comprises the lexical categories for adjectives (ADJ[1-12]) and ordinals (TW2), as well as the phrasal AP. (For evidence that the cardinals are nominal and the ordinals adjectival, see Van Eynde (2006).) V comprises the lexical categories for infinitives (WW[4-6]), past participles (WW[7-9]) and present participles (WW[10-12]), as well as the categories for bare infinitival phrases (INF), *te*-infinitives (TI), *om te*-infinitives (OTI), past participial phrases (PPART) and present participial phrases (PPRES). P comprises the lexical VZ[1-3] and the phrasal PP. The rest class comprises the lexical categories for finite verbs, adverbs, conjunctions, articles and interjections, as well as the phrasal categories for clauses, coordinate phrases and multi-word units.

	N	A	V	P	Rest	Sum	
zijn	2859	2759	327	463	785	7193	be
blijven	15	44	2	5	8	74	remain
lijken	10	19	1	0	10	40	seem
blijken	3	1	0	0	0	4	appear
schijnen	1	0	0	0	2	3	seem
heten	10	0	2	0	9	21	be called
er uit zien	3	16	1	0	7	27	look
worden	106	149	3	(1)	18	277	become
staan	(2)	52	2	13	8	77	stand
zitten	0	21	5	13	9	48	sit
liggen	(1)	22	2	5	7	37	lie
hangen	0	3	1	1	2	7	hang
kunnen	(1)	7	0	0	7	15	can
moeten	0	1	0	3	0	4	must
mogen	0	1	0	0	0	1	may
willen	0	1	0	2	2	5	want
hoeven	0	1	0	0	0	1	need
zien	0	2	0	0	0	2	look
klinken	0	17	0	0	6	23	sound
ruiken	0	1	0	0	0	1	smell
smaken	0	2	0	0	0	2	taste
aanvoelen	0	1	0	0	0	1	feel
overkomen	0	3	0	0	1	4	come across
voorkomen	0	1	0	0	0	1	appear
gaan	(1)	57	6	28	8	100	go
komen	0	14	0	29	2	45	come
(ge)raken	(1)	9	4	9	0	23	get
lopen	0	11	0	0	3	14	run
vallen	0	3	0	9	0	12	fall
doen	0	21	0	1	4	26	do
zich gedragen	0	2	0	0	0	2	behave
zich voordoen	0	1	0	0	0	1	pretend

Table 5: The category of the predicative complements in the sample

extent those which are treated as the prototypical copular verbs in Haeseryn et al. (1997), see section 2.

To avoid misunderstandings, it is worth stressing that the criterion is defined in terms of compatibility and not in terms of occurrence in the sample. There are two reasons for that. First, not all combinations which are well-formed can be expected to occur in the sample. *Lijken*, for instance, is compatible with a prepositional predicate, as in *dat lijkt in orde* 'that seems in order', but the sample just happens to lack any instances of it. Second, the sample contains dysfluencies and annotation errors, so that verbs which are not compatible with a nominal predicate can nonetheless be found with a nominal PREDC sister in the sample. *Liggen* in (21a) and *staan* in (21b), for instance, both have a nominal PREDC sister in the treebank, but the former (*de een na de ander*) is an adjunct, rather than a complement, and the latter (*rond de zestig zeventig*) is a PP.

(21) a. Als ik zie dat die dat die eigenlijk die kunststeden eigenlijk
when I see that those that those actually those art-cities actually
de één na de ander liggen hé. (sfv400295-221)
the one after the other lie uh

b. Mja rond de zestig zeventig staat die nu. (sfv400327-76)
yeah around the sixty seventy stands that now

A borderline case is *(ge)raken*. It is considered incompatible with nominal predicates by many speakers and the combination is indeed absent from the Europarl treebank (36.252.049 words), but it occasionally occurs in spontaneous speech and informal discourse. The sample contains one instance (*strop geraken* 'get stuck') and a Google search on April 22 2009 yielded a dozen of hits, including:[14]

(22) Gij zult nog burgemeester geraken als ge zo voortdoet.
you will still mayor get if you so continue
'You'll end up a mayor if you continue like that.'

Besides the constraint on the compatibility with nominal predicates, which neatly differentiates the prototypical copular verbs from the other predicate selectors, there are some other constraints, such as the incompatibility of *worden* 'become' with prepositional predicates. The sample, admittedly, contains one occurrence, but this is due to an annotation error: The PP *in Nederland* in (23) is treated as a dependent of *geworden*, whereas it is in fact a dependent of the demonstrative pronoun *dat*.

(23) Maar 't is uiteindelijk dat in Nederland geworden. (sfv901001-66)
But it is in-the-end that in the-Netherlands become
'But it has in the end become that in the Netherlands'.

The relevance of these constraints for NLP is obvious. For example, if one of the verbs in the lower part of the table is combined with a nominal complement,

[14] I thank Vincent Vandeghinste for the Europarl search and Geert Adriaens for the Google search.

NOMINATIVE	PREDC	OBJ1		ACCUSATIVE	PREDC	OBJ1	
ik	0	0	I	*mij, me*	0	431	me
wij, we	1	(1)	we	*ons*	0	239	us
jij	0	0	you	*jou*	0	30	you
hij, ie	2	0	he	*hem*	2	244	him
zij	0	0	she	*haar*	0	138	her
zij	0	0	they	*hen, hun*	0	94	them
Sum	3	(1)		Sum	2	1176	

Table 6: The case value of the pronominal complements in the sample

then we know that it must be an object complement, rather than a predicative one. Similarly, if *worden* is combined with a PP, then we know that it cannot be a predicative complement.

4.2 Morpho-syntactic selection

The difference between object selectors and predicate selectors is further underlined by the fact that the former impose constraints on the morpho-syntactic form of their complements whereas the latter do not.

Selectors of **nominal** objects require their complement to have some specific non-nominative case. The German *kennen* 'know', for instance, requires an accusative object and *helfen* 'help' a dative one. Predicate selectors, by contrast, do not impose any constraints on the case of their complement, so that it may also be nominative. Checking this in the sample is complicated by the fact that the distinction between nominative and accusative is systematically neutralized in Dutch, but it is not impossible, since some of the pronouns have separate forms for both cases. When their numbers are compared for PREDC and OBJ1 positions, the difference shows, see Table 6. While we find a low but nearly equal number for nominative and accusative forms in PREDCs, the high number of accusative forms in OBJ1s clearly contrasts with the virtual absence of nominative forms.[15]

Because of the lack of a case constraint, it is in principle possible that the choice of the case value is semantically significant. This is attested in Russian, where the contrast between nominative and instrumental case for predicate nominals corresponds with an aspectual distinction (Dalrymple et al. 2004, 192). Alternatively, it is possible to constrain the case value of the predicate in another way. In German and in Latin, for instance, the predicate nominals are required to show case agreement with their target. The predicate nominal in (24) must be nominative, just like the subject.[16]

[15]The one nominative form in object position is an instance of metalinguistic use: *Dat kwam omdat hij wij had gezegd.* (sfv800845-10) 'that's because he had said 'we".

[16]In Latin, but not in German, case agreement is also required for the adjectival predicates.

	PREDC			PREDC	
als	68	as	*beneden*	3	below
binnen	6	inside	*zonder*	3	without
buiten	4	outside	*naast*	2	next to
per	4	per	*boven*	1	above

Table 7: Prepositional complements

(24) Dieser Mann ist mein/*meinen Bruder.
This.NOM man is my.NOM/*my.ACC brother
'This man is my brother.'

Similar remarks apply to the **prepositional** complements. The selectors of prepositional objects canonically require their PP complements to be introduced by some specific preposition: *Wachten* 'wait', for instance, requires the presence of *op* and *zorgen* 'care' requires *voor*. The selectors of prepositional predicates, by contrast, do not impose any constraints of this kind. As a consequence, we find a greater variety of prepositions in predicative complements. The prepositions in Table 7, for instance, occur at least once in a PREDC, but not in any of PC, nor in OBJ1 or OBJ2. Another consequence is that the choice of the preposition may be semantically significant in predicates, which indeed it is. The preposition *voor* in *Hij is voor kernenergie* 'he is in favor of nuclear energy', for instance, semantically contrasts with *tegen* 'against'.

Also this information can be put to good use for NLP. Nominal complements which are nominative, for instance, must be predicative, and the same holds for prepositional complements which are introduced by one of the prepositions in Table 7.

4.3 Summing up

What is common to the two disambiguating factors is the fact that the object selectors impose stricter constraints on their complements than the predicate selectors. This is further underlined by a third factor: object selectors impose tighter constraints on the degree of saturation of their complements. The absence of a determiner in a singular count nominal **object**, for instance, tends to be unusual or even impossible, whereas it is unexceptional in a singular count nominal **predicate**, as illustrated by the contrast between **(een) leraar aanstellen* 'appoint *(a) teacher' and *(een) leraar worden* 'become (a) teacher'. This can be verified in the sample, albeit with greater difficulty, since the treebank does not contain information about the mass/count distinction.

5 Conclusion

Distinguishing the predicative complements from other types of complements is a matter of standard practice in linguistic theory and grammar writing. In spite

of that, the currently available treatments and descriptions are remarkably vague. They only mention the prototypical predicate selectors, assuming that the reader will be able to extrapolate, and they contain little information about how the predicate selecting uses of the relevant verbs can be differentiated from their other uses. The aim of this paper was to demonstrate how the wealth of data which treebanks contain can be exploited in order to arrive at a better and more precise understanding of what is is that differentiates the predicative complements from the other complements. The relevance of this knowledge for NLP purposes has been highlighted at the appropriate places.

References

Bouma, G. (2004), A corpus investigation of PP-fronting in Dutch, *in* Decadt, B., V. Hoste, and G. De Pauw, editors, *Computational Linguistics in the Netherlands 2003*, Antwerp Papers in Linguistics, pp. 15–29.

Dalrymple, M., H. Dyvik, and T. Holloway King (2004), Copular complements: closed or open?, *in* Butt, M. and T. Holloway King, editors, *Proceedings of the LFG04 Conference*, CSLI Publications, Stanford, pp. 188–198.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij, and M.C. van den Toorn (1997), *Algemene Nederlandse Spraakkunst*, Nijhoff and Wolters Plantyn.

Hinrichs, E. and H. Telljohan (2009), Constructing a valence lexicon for a treebank of German, *in* Van Eynde, F., A. Frank, K. De Smedt, and G. van Noord, editors, *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories TLT7*, LOT, Utrecht, pp. 41–52.

Hoekstra, H., M. Moortgat, B. Renmans, M. Schouppe, I. Schuurman, and T. van der Wouden (2003), CGN syntactische annotatie, Utrecht/Leuven.

Kupsc, A. and A. Abeillé (2008), Growing treelex, *in* Gelbukh, A., editor, *CICLing*, Springer Verlag, Berlin, pp. 28–39.

Oostdijk, N. (2000), Building a corpus of spoken Dutch, *in* Monachesi, P., editor, *Computational Linguistics in the Netherlands 1999*, Utrecht University, Utrecht.

Pollard, C. and I. Sag (1994), *Head-driven Phrase Structure Grammar*, CSLI Publications and University of Chicago Press, Stanford/Chicago.

Sarkar, A. and D. Zeman (2000), Automatic extraction of subcategorization frames for Czech, *COLING 2000*, Morgan Kaufmann, Saarbrücken, pp. 691–697.

van den Toorn, M.C. (1969), De bepaling van gesteldheid, *De nieuwe taalgids* **62**, pp. 34–40.

Van der Wouden, T., I. Schuurman, M. Schouppe, and H. Hoekstra (2003), Harvesting Dutch trees: syntactic properties of spoken Dutch, *in* Gaustad, T., editor, *Computational Linguistics in the Netherlands 2002*, Rodopi, pp. 129–141.

Van Eynde, F. (2006), NP-internal agreement and the structure of the noun phrase, *Journal of Linguistics* **42**, pp. 139–186.

Van Eynde, F. (2008), Predicate complements, *in* Müller, S., editor, *On-Line Proceedings of HPSG 2008*, CSLI Publications, Stanford University, pp. 253–273.