Available online at www.sciencedirect.com

ScienceDirect

Lingua 178 (2016) 104–126

Lingua

www.elsevier.com/locate/lingua

Number agreement in copular constructions: A treebank-based investigation

Frank Van Eynde *, Liesbeth Augustinus, Vincent Vandeghinste

Center for Computational Linguistics, University of Leuven, Belgium

Received 7 October 2014; received in revised form 2 February 2016; accepted 2 February 2016
Available online 27 March 2016

Abstract

This paper has both a theoretical and a methodological objective. The theoretical one concerns the modeling of number agreement in copular constructions. For that purpose it adopts the distinction, familiar from Head-driven Phrase Structure Grammar, between morpho-syntactic agreement (also known as concord) and index agreement. The methodological objective concerns the demonstration of how treebanks can be exploited in order to guide the formulation of relevant generalizations. For that purpose we crucially rely on tools and resources that have recently been developed in the framework of the Dutch-Flemish STEVIN program (2004–2011) and the European CLARIN infrastructure.

Keywords: Copular construction; Number agreement; Predicate nominal; Treebanks; Concord; Index sharing; Head-driven Phrase Structure Grammar; Distributive; Collective

1. Introduction

This paper focusses on constructions which consist of a subject, a copular verb and a predicate nominal. In such constructions there is not only number agreement between the subject and the finite verb, but also between the subject and the predicate nominal, as illustrated in (1).

(1) a. His brother is an engineer.
b. * His brother is engineers.
(2) a. His brothers are both engineers.
b. * His brothers are both an engineer.

Mismatches, however, are not excluded. The sentences in (3), for instance, are well-formed.[1]

(3) a. I am best friends with the president of Finland.
b. His brothers are a danger on the road.

* Corresponding author. Tel.: +32 16 325084; fax: +32 16 325098.
E-mail address: frank.vaneynde@ccl.kuleuven.be (F. Van Eynde).
[1] (3a) is quoted from Fillmore et al. (2012:351).

http://dx.doi.org/10.1016/j.lingua.2016.02.001

Table 1
Contents of the LASSY treebank.

Label	Contents	# sentence	# word
wr-p-p	Books, brochures, newspapers, reports, periodicals and magazines, proceedings, legal texts, policy documents, surveys, guides and manuals	17,691	281,424
wr-p-e	E-magazines, newsletters, web sites, teletext pages	14,420	232,631
ws-u	Auto cues, news scripts, texts for the visually impaired	14,032	184,611
dpc	Dutch Parallel Corpus	11,716	193,029
Wikipedia	Dutch Wikipedia pages	7,341	83,360
Sum		65,200	975,055

The challenge for a treatment of these data is to make it sufficiently restrictive to enforce agreement when it is required and sufficiently flexible to allow for mismatches. To pave the way for such a treatment we adopt a corpus based approach. Making use of a Dutch treebank, to be presented in section 2, we extract the relevant agreement data in ways that are described in section 3. A quantitative analysis of the data unambiguously shows the agreement effect, but it also reveals that mismatches as those in (3) occur in sufficiently high frequency to justify a more detailed investigation. This is undertaken in section 4, which presents a typology of mismatches. Building on that typology we present a unified formal treatment of the data in section 5. It is developed in the framework of Head-driven Phrase Structure Grammar (HPSG). The conclusions are summed up in section 6.

2. The LASSY treebank

LASSY is a treebank of written Dutch. It was constructed in the framework of the STEVIN program (Spyns and Odijk, 2013) and is described in Van Noord et al. (2013). Table 1 provides a survey of the types of texts that the treebank contains and of their size in terms of sentences and words.[2]

The texts are divided in sentences and each sentence has a unique identifier, as in (4).

(4) De slachtoffers zijn volgens de verkeerspolitie vermoedelijk Nederlanders.
the victims are according the traffic.police probably Dutch.ones
'The victims are probably Dutch according to the traffic police.'
(ws-u-e-a-0000000205.p.18.s.2)

Each sentence is assigned a tree that contains information about syntactic categories and dependencies, in accordance with the annotation guidelines in Hoekstra et al. (2003). The tree for (4), for instance, is given in Fig. 1.

The italicized word tokens at the bottom of the tree are assigned a lemma and a lexical category. The names of the lexical categories are abbreviations of Dutch terms: 'ww' is short for 'werkwoord' (verb), 'vz' for 'voorzetsel' (preposition), 'lid' for 'lidwoord' (article), and so on. Phrases have at least two daughters and are assigned a phrasal category, such as 'np' or 'pp'. Both the lexical and the phrasal nodes also contain a dependency label, such as 'h(ea)d' or 'mod(ifier)'.[3] Notice that the trees are relatively flat: The subject, the predicative complement and the two modifiers are all sisters of the verbal head in Fig. 1.

The lexical categories are abbreviations of more detailed part-of-speech tags. These tags contain information about various morpho-syntactic distinctions, in accordance with the annotation guidelines in Van Eynde (2003). One of the distinctions concerns morpho-syntactic number. More specifically, the nouns and the pronouns have a feature, called 'getal' (number), whose value is either 'enkelvoud' (singular) or 'meervoud' (plural). For the pronouns, the value may be left underspecified. An example is the demonstrative *die* 'that/those', which is compatible with both singular and plural verbs.

(5) a. Die is echt gevaarlijk.
that is really dangerous
'That one is really dangerous.'
b. Die zijn echt gevaarlijk.
that are really dangerous
'Those are really dangerous.'

[2] These are the numbers for LASSY Small, i.e. the part of the treebank for which the output of the ALPINO parser was manually checked and, if necessary, corrected. There is also LASSY Large, in which the output of the parser is not manually checked. For a description of the ALPINO parser, see Van Noord (2006).

[3] The immediate daughters of the top node, which include the sentence final punctuation, are assigned the vacuous dependency label '-', see Fig. 1.

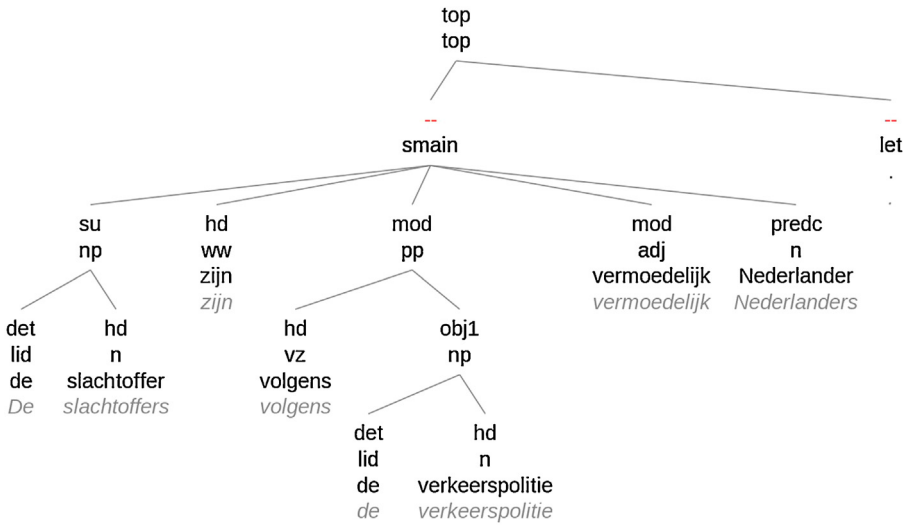

Fig. 1. Syntactic representation of (4).

Table 2
The (pro)nouns and their number in the LASSY treebank.

	Noun	Pronoun	Sum	%
Singular	188,297	25,900	214,197	71.30
Plural	58,458	8,265	66,723	22.21
Underspecified	–	19,486	19,486	6.49
Sum	246,755	53,651	300,406	

Table 2 provides some quantitative data about the nouns, the pronouns and their morpho-syntactic number. Notice that they jointly account for nearly one third of the words in the treebank (300,406/975,055).

The quality of the annotations has been assessed by the Danish Centre for Language Technology. Extrapolating from a sample of 500 randomly selected sentences they estimate that the dependency labels are all correctly assigned in 97.8% of the sentences and that 98.63% of the words are correctly tagged (Jongejan et al., 2011).

3. Quering the treebank

This section describes how the relevant data are extracted from the treebank (section 3.1) and how they are checked for accuracy and relevance (section 3.2). The results are summed up in section 3.3.

3.1. Automatic extraction

The information which is needed to check the number agreement in copular constructions is extracted by means of queries that are expressed in terms of XPath notation.[4] (6), for instance, retrieves the combinations of a verbal head, a subject and a predicative complement in which both the subject and the predicative complement are marked for number (@getal). It yields 164 hits.

[4] See http://www.w3.org/TR/xpath/ for a description of the notation. The queries can be expressed directly in XPath notation or they can be derived automatically from a sample sentence. For the latter, we use GrETEL. This search tool invokes the ALPINO parser for an analysis of the given example, it allows the user to identify those aspects of the parse that are considered relevant for the search, and it automatically translates the resulting pattern in an XPath query, see Augustinus et al. (2012) and http://gretel.ccl.kuleuven.be. GrETEL was developed in the framework of the Flemish CLARIN project Nederbooms (2010–2012).

Table 3
Results of the automatic extraction.

SU–PREDC	sg-sg	sg-pl	sg-und	pl-sg	pl-pl	und-x	Sum
X–X	130	8	2	12	7	5	164
XP–X	79	11	0	19	18	2	129
X–XP	1640	142	11	53	46	23	1915
XP–XP	1272	22	4	137	90	2	1527
Sum	3121	183	17	221	161	32	3735

(6) //node[
node[@rel=''hd'' and @pt=''ww''] and
node[@rel=''su'' and @getal] and
node[@rel=''predc'' and @getal]]

The mutual order of the nodes does not matter: (6) retrieves instances of any of the six possible orders of the three daughters. What it does not retrieve, though, are combinations in which the subject or the predicative complement is phrasal. This is due to the fact that the number feature is only assigned to lexical categories. To retrieve the relevant phrasal categories as well we need queries in which the subject or the predicative complement is required to contain a head daughter that has the number feature. (7), for instance, retrieves the combinations in which both are phrasal. It yields 1527 hits.[5]

(7) //node[
node[@rel=''hd'' and @pt=''ww''] and
node[@rel=''su'' and node[@rel=''hd'' and @getal]] and
node[@rel=''predc'' and node[@rel=''hd'' and @getal]]]

Predictably, there are also combinations in which only the subject is phrasal (129 hits) and in which only the predicative complement is phrasal (1915 hits).

In a second step, we add specific values for the number features of the subject and the predicative complement. (8), for instance, retrieves the combinations in which they are both lexical and singular. It yields 130 hits.[6]

(8) //node[
node[@rel=''hd'' and @pt=''ww''] and
node[@rel=''su'' and @getal=''ev''] and
node[@rel=''predc'' and @getal=''ev'']]

Repeating such queries for the other combinations, we get the data that are displayed in Table 3.

The first column specifies whether the subject and the predicative complement are lexical (X) or phrasal (XP). The next three columns provide information about the combinations with a singular subject, differentiating the combinations in which also the predicate nominal is singular (sg-sg) from the combinations in which it is plural (sg-pl) or underspecified (sg-und). The next columns do the same for the combinations with a plural subject.[7] The penultimate column concerns the combinations with a subject that is underspecified for number (und-x). We do not make any finer-grained distinctions in that category, since they are irrelevant for checking agreement. For our purpose it is mainly the figures in the bottom line that matter, but we abstain at this stage from discussing them, since they first need to be checked for accuracy and relevance.

3.2. Manual inspection

Manually inspecting a sample of 3735 instances is a very time consuming task. To keep it manageable we have set the 'sg-sg' combinations aside and focussed on the other combinations which jointly account for 614 instances. For each of

[5] The query does not retrieve coordinate NPs, since they do not have a head daughter.
[6] ''ev'' is short for ''enkelvoud'' (singular).
[7] There are no instances of the 'pl-und' combination in the treebank.

them we have checked whether they are genuine instances of the said combination. If not, they are classified as annotation errors or irrelevant hits.

3.2.1. Annotation errors

We found 29 annotation errors.[8] They can be divided in three groups. The largest one concerns the assignment of an erroneous number value (12 hits). The predicative complement in (9), for instance, is not the plural counterpart of *naam* 'name', but a singular proper noun, and the subject in (10) is not a noun with a plural affix, but with the feminizing affix *-es*.

(9) De hoofdstad van Wallonië is Namen.
the capital of Wallonia is Namur
'The capital of Wallonia is Namur.' (wiki-135.p.36.s.1)

(10) De benedictines Joan Chittister is de enige spreekster op het congres die wat
the benedictinesse Joan Chittister is the only speaker at the congress who something
te vrezen heeft van haar kerk.
to fear has of her church
'The benedictinesse Joan Chittister is the only speaker at the congress who has something to fear of her church.'
(wr-p-p-h-0000000048.p.8.s.1)

The second group concerns the assignment of an inappropriate category label (9 hits). In (11), for instance, the predicative complement is not a plural noun, but an adjective, and in (12) the subject is not a plural common noun, but a nominal infinitive.

(11) De ambtenaar is lui.
the civil.servant is lazy
'The civil servant is lazy.' (dpc-vla-001161-nl-sen.p.117.s.2)

(12) Reizen met de auto was tot ruim een jaar geleden een waagstuk.
travel with the car was till good a year ago a risk
'Traveling by car was risky till a good year ago.'
(ws-u-e-a-0000000010.p.24.s.3)

The third group concerns the assignment of an inappropriate dependency label (8 hits). A relevant example is (13).

(13) ... zoals de voorbije weken het geval was
... as the past weeks the case was
'... as was the case during the past weeks' (wr-p-p-i-0000000204.p.2.s.2)

The plural NP *de voorbije weken* 'the past weeks' is not the subject, but a temporal modifier.

3.2.2. Irrelevant hits

34 of the hits turned out to be irrelevant. Most of them are combinations with an object-oriented predicative complement (28 hits). A relevant example is the indefinite predicate nominal in (14).

(14) Ze noemen het wel een stap in de goede richting.
they call it POL a step in the right direction
'They call it a step in the right direction.' (ws-u-e-a-0000000022.p.33.s.1)

Such combinations should be ignored, since object-oriented predicative complements are expected to show number agreement with the direct object, rather than with the subject. They are easy to retrieve in the sample since the verbs which combine with an object-oriented predicate nominal form a small class, including *noemen* 'call', *vinden* 'consider', *maken* 'make' and *achten* 'deem'.

[8] It is possible that more recent releases of the treebank no longer contain (some of) these annotation errors, since they have been reported to its chief developer.

A second group concerns combinations with an autoreferential predicate nominal (4 hits). A relevant example is (15).

(15) Het thema dit jaar is ‘‘Steden’’.
the theme this year is ‘‘Cities’’
‘The theme this year is ‘‘Cities’’.’ (wr-p-e-c-0000000004.p.15.s.6)

We do not count this as a ‘sg-pl’ mismatch, since autoreferential nominals are singular for the purpose of agreement, not only in predicative position, but also in the subject position of a finite clause, as shown in (16).

(16) ‘‘Steden’’ is/*zijn een geschikte titel voor dit boek.
‘‘Cities’’ is/*are an appropriate title for this book
‘‘‘Cities’’ is/*are an appropriate title for this book.’

Completing the list of irrelevant hits are the examples in (17–18).

(17) Ook voor ... de politieke discussies ... zijn ze groot belang.
Also for ... the political discussions ... are they great importance
‘They are also (of) great importance for the political discussions.’
(wr-p-p-j-0000000013.p.39.s.2)

(18) Andere opvallende klemtonen in het nieuwe plan zijn:
Other conspicuous accents in the new plan are:
- Verankering van het plan in het selectiebeleid ...
- Anchoring of the plan in the selection.policy ...
‘Other conspicuous accents in the new plan are: - Anchoring of the plan in the selection policy...’
(wr-p-e-j-0000000009.p.3.s.1)

(17) is an ill-formed sentence: The predicate nominal *groot belang* ‘great importance’ should be introduced by the preposition *van* ‘of’. When that is repaired, there is no mismatch, since predicative PPs are not expected to show number agreement with the subject. (18) ends up in the list of mismatches because it is part of an enumeration whose other members are treated as separate sentences, rather than as conjuncts of the same PREDC. When that is repaired, there is no mismatch.

3.2.3. Summing up

The result of eliminating the annotation errors and irrelevant hits is provided in Table 4.

The numbers in the first row are identical to those in the bottom line of Table 3 (the ‘sg-sg’ combination is left out here). The figures for the annotation errors are in the second row and those for the irrelevant hits in the third row. They jointly account for 63 hits, which on a total of 614 instances amounts to 10.26%. Assuming that a similar proportion of the ‘sg-sg’ hits concerns annotation errors or irrelevant hits, it can be estimated that approximately 320 of those hits are to be subtracted, yielding a sum of approximately 2801 genuine ‘sg-sg’ hits.

3.3. Results

Adding the estimate for the ‘sg-sg’ combination and leaving out the combinations with an underspecified number value, we get the data in Table 5.

The matches (sg-sg and pl-pl) amount to 89.48% of the total number of combinations (2959/3307) and, hence, clearly outnumber the mismatches. The correlation between the number values of the subject and the predicate nominal is

Table 4
Results after manual inspection.

SU–PREDC	sg-pl	sg-und	pl-sg	pl-pl	und-x	Sum
Extracted	183	17	221	161	32	614
Annotation errors	−11	−2	−15	−1		−29
Irrelevant hits	−7		−23	−2	−2	−34
Result	165	15	183	158	30	551

Table 5
Matches vs. mismatches.

	PREDC-SG	PREDC-PL	Sum
SU-SG	2801	165	2966
SU-PL	183	158	341
Sum	2984	323	3307

Table 6
The predicate nominals and their number in the LASSY treebank.

PREDC	X	XP	Sum	%
Singular	474	4758	5232	90.19
Plural	75	470	545	9.39
Underspecified	7	17	24	0.41
Sum	556	5245	5801	

Table 7
Four types of mismatches.

SU–PREDC	sg verb	pl verb	Sum
sg-pl	12	153	165
pl-sg	7	176	183
Sum	19	329	348

confirmed by a χ^2 test. It shows that the null hypothesis, namely that the number values are unrelated, can be rejected with a 99% certainty.[9] There is, hence, a clear agreement effect.

Further confirmation is provided by a comparison with the proportion of singular and plural predicate nominals in the treebank as a whole. Employing queries, such as those in (19–20), which retrieve the lexical and phrasal singular predicate nominals respectively, we get the data in Table 6.[10]

(19) //node[@rel=''predc'' and @getal=''ev'']
(20) //node[@rel=''predc'' and node[@rel=''hd'' and @getal=''ev'']]

If there were no agreement effect, we would expect the frequency of singular predicate nominals in clauses with a plural subject to be around 90%, but it is 53.67% (=183/341). Similarly, we would expect the frequency of plural predicate nominals in clauses with a singular subject to be around 10%, but it is 5.56% (=165/2966).

At the same time, the amount of mismatches (10.52%) is too high to be dismissed as unworthy of attention.

4. A typology of mismatches

For an investigation of the mismatches we also take the number value of the finite verb into account. We, hence, differentiate the 'sg-pl' mismatches with a singular finite verb from those with a plural finite verb, and do the same for the 'pl-sg' mismatches. This yields the four-fold classification in Table 7.

For each of the four types this section provides examples and some informal discussion. We first discuss the combinations in which the finite verb shows agreement with the subject (sections 4.1 and 4.2), and then those in which it does not (sections 4.3 and 4.4). A summary is provided in section 4.5.

[9] The χ^2 statistic is 576.8551. The p value is 0. The result is significant at $p < 0.01$.
[10] The total number of predicate nominals in this table is higher than in Table 3, since the queries in (19–20) retrieve predicate nominals, rather than triples of a verb, a subject and a predicate nominal.

4.1. Plural verb and plural subject vs. singular PREDC

This is the type of mismatch that is exemplified by the English (3b), repeated in (21).

(21) His brothers are a danger on the road.

With 176 instances, it is the most common type of mismatch in the treebank. Some examples are given in (22–23).

(22) Beide aftredende bestuurders blijven wel aandeelhouder.
both resigning directors remain POL shareholder
'Both resigning directors do remain shareholder.'
(wr-p-e-i-0000049645.p.1.s.68.2)

(23) De Tsjetsjeense strijders zijn een relatief kleine groep.
the Chechen warriors are a relatively small group
'The Chechen warriors are a relatively small group.'
(ws-u-e-a-0000000244.p.3.s.4)

(22) has a distributive interpretation: It entails that both of the resigning directors remain shareholders. (29), by contrast, has a collective interpretation: It does not entail that there are as many groups as there are Chechen warriors, but rather that there is one group that consists of a small number of Chechen warriors. Since the distinction will turn out to be important for the formal treatment in section 5, we submit it to a closer look.

First, we propose a test to differentiate the combinations with a distributive interpretation from those with a collective interpretation: If the substitution of the predicate nominal by its plural counterpart does not entail a difference in meaning, then the combination has a distributive interpretation. This is vividly illustrated in (24).

(24) Dat betekent niet dat de initiatiefnemers nu ineens managers zijn.
that means not that the initiators now suddenly managers are.
Ze zijn en blijven vooral boer.
they are and remain chiefly farmer
'That does not mean that the initiators are now all of a sudden managers. They are and remain in the first place farmers.'
(ws-u-e-a-0000000217.p.26.s.2)

The predicate nominal in the subordinate clause of the first sentence is the plural *managers* 'managers', while the one in the second sentence is the singular *boer* 'farmer', but both clauses have the same kind of distributive interpretation, and this does not change if *managers* is replaced by the singular *manager* or if *boer* is replaced by the plural *boeren*. In combinations with a collective interpretation this equivalence of singular and plural forms does not hold: If the predicate nominal in (23) is replaced by its plural counterpart, the sentence no longer talks about one group of warriors, but about several such groups.

Second, we explore the relation between the properties of the predicate nominal and the interpretation of the clause as a whole. For that purpose we classify the predicate nominals in terms of two dimensions. One concerns the internal structure of the predicate nominal, differentiating the bare predicate nominals from the definite and indefinite ones. The other concerns the position of the predicate nominal in the sentence, differentiating those which follow the subject, as in (23–24), from those which precede it, as in (25).

(25) Het grootste probleem tijdens de wedstrijden zijn de spreekkoren.
the main problem during the matches are the speech.choirs
'The main problem during the matches are the choirs.'
(ws-u-e-a-0000000218.p.7.s.3)

Table 8 provides a survey.

For the discussion we first focus on the combinations which show the canonical word order, i.e. with the subject before the predicate nominal.

Table 8
Survey of the type 1 mismatches.

PREDC	Bare	Definite	Indefinite	Sum
Canonical	40	50	69	159
Topicalized	6	5	6	17
Sum	46	55	75	176

4.1.1. Bare predicate nominals

The bare predicate nominals lack a determiner. This absence is not limited to nominals which are headed by a mass noun. It also affects nominals which are headed by a count noun, such as *aandeelhouder* 'shareholder' or *boer* 'farmer'.

4.1.1.1. The combinations with a bare predicate nominal canonically have a **distributive** interpretation. This has already been demonstrated for (22) and (24). Another example is (26).

(26) Overigens zullen de drempels niet gelden voor werknemers uit Malta en Cyprus.
By.the.way will the tresholds not apply for workers from Malta and Cyprus.
Die eilanden worden per 1 mei óók EU-lidstaat.
Those islands become on 1 May also EU-member.state
'The tresholds will not apply to workers from Malta and Cyprus. Those islands become EU member states as well on May 1st.'
(ws-u-e-a-0000000043.p.9.s.6)

The anaphoric *die eilanden* 'those islands' denotes two islands which both become a member state of the European Union. Notice the semantic equivalence of the singular and the plural forms of the predicate nominal: Replacing the singular *EU-lidstaat* 'EU member state' with its plural counterpart does not affect the meaning of (26).

4.1.1.2. While the assignment of the distributive interpretation is always possible in combinations with a bare predicate nominal, there are cases in which the **collective** interpretation is more plausible, as in (27).

(27) Vorig jaar werden we kampioen met 84 punten.
last year became we champion with 84 points
'Last year we became the champion with 84 points.'
(dpc-rou-000358-nl-sen.p.3.s.15)

This sentence can be assigned a distributive interpretation, namely if the subject *we* 'we' is understood to denote at least two individuals or teams which both became champion with 84 points, for instance, if *we* stands for the union of a Spanish and a German football team which won the Spanish, resp. German championship, each with 84 points. While this is not impossible, the collective interpretation, in which *we* stands for one team that became champion, is clearly more plausible.

Overall, though, (27) is the exception: Of the 40 instances in the sample, it is the only one for which the collective interpretation is preferable.

4.1.2. Definite predicate nominals

Most of the definite predicate nominals in the sample are introduced by the definite article (*het* 'the' in neuter nominals and *de* 'the' in non-neuter nominals). Besides, there are two with a possessive determiner, one with a demonstrative determiner and one pronoun.

4.1.2.1. The combinations with a definite predicate nominal have a **distributive** interpretation when the predicate nominal denotes a role, as in (28).

(28) Volgens sommige bronnen werden minstens 156 mensen hiervan het slachtoffer.
following certain sources became at.least 156 people here.of the victim
'According to certain sources at least 156 people became the victim of this.' (wr-p-e-i-0000004745.p.5.s.36)

(28) is about 156 people who are all victims. Typical role denoting NPs are *het slachtoffer* 'the victim', *de dupe* 'the victim' and *het doelwit* 'the target'. They are occasionally used without determiner, as in (29).

(29) Ook Wolderse jongens stonden toen in die eerste linies,
also Wolderian boys stood then in those first lines,
ook zij werden slachtoffer en voelden zich verraden.
also they became victim and felt REFL betrayed
'Boys from Wolder also stood in those first lines, they also became victims and felt betrayed.' (wr-p-e-i-0000050381.p.1.s.430)

This chimes well with the fact that combinations with a bare predicate nominal canonically have a distributive interpretation.

4.1.2.2. The combinations with a definite predicate nominal have a **collective** interpretation when the predicate nominal is understood to denote a unique referent, as in (30).

(30) De Leien (Frankrijklei, Italiëlei, Amerikalei, Britselei) zijn de belangrijkste verkeersader binnen Antwerpen.
the Leien (Frankrijklei, Italiëlei, Amerikalei, Britselei) are the important.SUP traffic.artery in Antwerp
'The Leien (Frankrijklei, Italiëlei, Amerikalei, Britselei) are the most important traffic artery in Antwerp.'
(wiki-11.p.54.s.1)

(30) is about one trajectory that is claimed to be the most important one in Antwerp and that trajectory is identified with the 'Leien' as a whole. If one wants to claim that each one of the four 'Leien' belongs to the most important trajectories of the city, one must use the plural counterpart, i.e. *de belangrijkste verkeersaders* 'the most important traffic arteries'. That the predicate nominal in (30) has a unique referent is underlined by the presence of the superlative modifier *belangrijkste* 'most important'. Other uniqueness markers are the modifier *enige* 'only' in (31) and the proper noun in the predicate nominal of (32).

(31) ... omdat in de plannen de roltrappen de enige vluchtweg uit de ondergrondse zijn.
... because in the plans the escalators the only escape.route from the underground are
'... because the escalators are the only way out from the underground.'
(ws-u-e-a-0000000200.p.15.s.2)

(32) De kernen Heukelom en Montenaken werden de gemeente Vroenhoven.
the nuclei Heukelom and Montenaken became the town Vroenhoven
'The nuclei Heukelom and Montenaken became the town Vroenhoven.'
(wr-p-e-i-0000050381.p.1.s.148)

The collective interpretation is also the only plausible one for the combination with the demonstrative determiner in (33).[11]

(33) Op de internationale wisselmarkten doen de twee toch al of ze dezelfde munt zijn.
on the international exchange.markets do the two anyway as if they the.same currency are
'On the international exchange markets the two behave anyway as if they were the same currency.'
(wr-p-p-i-0000000058.p.3.s.3)

The assignment of a distributive interpretation to (33) would be absurd: Saying of one single currency that it is the same does not make sense.

4.1.3. Indefinite predicate nominals

Most of the indefinite predicate nominals in the sample are introduced by the indefinite article *een* 'a(n)' or its negative counterpart *geen* 'no'. Besides, there is one which is introduced by *voldoende* 'sufficient' and there are three indefinite pronouns.

[11] Notice that *dezelfde* 'the same' is treated as a single word in Dutch orthography.

4.1.3.1. Some of the combinations with an indefinite predicate nominal have a **distributive** interpretation. A relevant example is (34).

(34) De Arabische staten die onder Brits bewind hadden gestaan werden veelal een monarchie.
the Arab states which under British rule had stood became mostly a monarchy
'The Arab states which had been under British rule, mostly became monarchies.'
(wr-p-e-i-0000015007.p.1.s.175)

This sentence does not mean that the Arab states which were part of the British empire now collectively constitute one monarchy, but rather that a number of such states have become monarchies. This is confirmed by the replacement test: If the indefinite is replaced by its plural counterpart (*monarchieën* 'monarchies') the resulting interpretation is the same. Of special relevance in this case is the presence of the frequency adjunct *veelal* 'mostly'. When it is omitted, the collective interpretation becomes more plausible.

Notice, though, that the presence of such an adjunct is not necessary to trigger a distributive interpretation. The combination *dat ze een goede moslim zijn* 'that they are a good muslim' in (35), for instance, has a distributive interpretation, even though it does not contain any element which triggers it explicitly.

(35) Zo zullen steeds minder jongemannen zichzelf in een volgende generatie ervan kunnen
so will ever fewer youngsters REFL in a next generation it.of can
overtuigen dat ze ''een goede moslim'' zijn als ze onschuldige medemensen afmaken.
convince that they ''a good muslim'' are if they innocent people kill
'Always fewer youngsters will be able to convince themselves that they are ''a good muslim''
if they kill innocent people.'
(dpc-ind-001636-nl-sen.p.19.s.4)

Notice, also here, the semantic equivalence of the singular predicate nominal with its plural counterpart *goede moslims* 'good muslims'.

An example with the negative *geen* 'no' is given in (36).

(36) De artsen die Millecam hebben behandeld waren overigens geen lid van de
the doctors who Millecam have treated were by-the-way no member of the
beroepsvereniging voor homeopaten.
syndicate of homeopaths
'The doctors who treated Millecam were not members of the syndicate of homeopaths.'
(ws-u-e-a-0000000049.p.18.s.2)

Notice that the non-negative counterpart of this predicate nominal is the bare singular *lid* 'member' and that replacement with the plural *geen leden* 'no members' yields a sentence that is semantically equivalent to (36).

4.1.3.2. Most of the combinations with an indefinite predicate nominal have a **collective** interpretation, not only when the predicate nominal is headed by an inherently collective noun, such as *groep* 'group' in (23), but also when it is headed by another kind of noun, as in (37).

(37) Politieke tegenstellingen zijn een wezenskenmerk van elke democratie.
political contrasts are a defining.feature of every democracy
'Political contrasts are a defining characteristic of every democracy.'
(dpc-kok-001320-nl-sen.p.6.s.2)

This sentence does not mean that each political contrast is a defining characteristic of democracy, but rather that the existence of such contrasts in general is a characteristic of democracy.

The collective interpretation is also preferable for combinations like (38).

(38) Omdat we geen nationale carrier zijn ...
because we no national carrier are ...
'Because we are not a national carrier.' (wr-p-p-i-0000000183.p.9.s.2)

Replacement with the plural *geen nationale carriers* 'no national carriers' yields a sentence with another meaning than that of (38). Notice also that the non-negative counterpart of the predicate nominal is not a bare singular but the indefinite *een nationale carrier* 'a national carrier'.

The collective interpretation also applies to the combination with *voldoende* 'sufficient'.

(39) Die bevindingen zijn voor de Inspectie voldoende aanleiding strafbare feiten te constateren ...
those considerations are for the Inspection sufficient reason punishable facts to establish ...
'Those considerations are sufficient reason for the Inspection to establish punishable facts...' (ws-u-e-a-0000000047.p.7.s.5)

What is claimed to be a sufficient reason in (39) is not each single one of the considerations, but rather the sum of them.

4.1.4. Topicalized predicate nominals

The topicalized predicate nominals in the sample include six bare nominals, five definite NPs and six indefinite NPs. (40–42) provide an example of each.

(40) Grote winnaar bij de verkiezingen van 18 mei 2003 waren de socialisten.
great winner at the elections of 18 May 2003 were the socialists
'The socialists were the great winners of the elections of May 18, 2003.'
(wr-p-e-h-0000000051.p.249.s.1)

(41) Het grootste probleem tijdens de wedstrijden zijn de spreekkoren.
the main problem during the matches are the speech.choirs
'The main problem during the matches are the choirs.'
(ws-u-e-a-0000000218.p.7.s.3)

(42) Een trekpleister in het dorp zijn de druipsteengrotten.
an attraction.pole in the village are the stalactite.caves
'A pole of attraction in the village are the caves with stalactites.'
(wiki-9515.p3.s.1)

These combinations all have a **collective** interpretation: (40) attributes the big victory to the socialists as a group, (41) is about one main problem, and (42) is about one pole of attraction.

The assignmment of a collective interpretation to (41) and (42) is not surprising, since the corresponding non-topicalized sentences (with the predicate nominal after the subject) also have a collective interpretation. It is remarkable, though, for the combination with the bare predicate nominal in (40), since combinations with a bare predicate nominal canonically have a distributive interpretation. This peculiarity correlates with another one: In the non-topicalized counterpart of (40) the predicate nominal must be introduced by the definite article.

(43) De socialisten waren de grote winnaar bij de verkiezingen van 2003.
the socialists were the great winner with the elections of 2003
'The socialists were the great winner of the 2003 elections.'

This makes the assignment of a collective interpretation to (40) more in line with the expectations, since the definite predicate nominal in (43) has a unique referent.

4.1.5. Summing up

The results of this survey can be summed up as follows:

- combinations with a bare predicate nominal canonically have a distributive interpretation;
- combinations with a definite predicate nominal have a distributive interpretation if the predicate nominal denotes a role, and a collective one if it has a unique referent;
- combinations with an indefinite predicate nominal may have a distributive interpretation, but the collective one is more common;
- combinations with a topicalized predicate nominal have a collective interpretation, no matter whether the predicate nominal is bare, definite or indefinite.

4.2. Singular verb and singular subject vs. plural PREDC

This is the type of mismatch that is exemplified by the English (3a), repeated in (44).

(44) I am best friends with the president of Finland.

It is rather uncommon in Dutch. The treebank, for instance, contains only twelve instances. Nine of them are combinations in which the predicate nominal contains a numeral, as in (45).

(45) Bij een vrouw is de grens veertien glazen.
with a woman is the limit fourteen glasses
'For a woman the limit is fourteen glasses.'
(wr-p-p-c-0000000048.txt-341)

This sentence clearly has a collective interpretation: There is one limit and it is identified with (the consumption of) fourteen glasses (of alcoholic beverages per week). Of the remaining combinations two have a predicate nominal which is headed by a *plurale tantum*, such as *activa* in (46).[12]

(46) Goud blijft de belangrijkste financiële activa van bijna alle centrale banken.
gold remains the important.SUP financial activa of nearly all central banks
'Gold remains the most important financial activa of nearly all central banks.' (wr-p-e-i-0000032165.p.5.s.255)

These also have a collective interpretation. Completing the survey is the combination with a topicalized predicate nominal in (47).

(47) De naamsveranderingen van de partijen was niet de enige wijziging.
the name.changes of the parties was not the only change
'It was not only the names of the parties that changed.'
(dpc-rou-000479-nl-sen.p.10.s.8)

Also here the most plausible interpretation is the collective one. In contrast then to the mismatches of type 1, these all have a collective interpretation.

4.3. Plural verb and plural PREDC *vs. singular subject*

This combination shows a double mismatch: It is not only the predicate nominal that has another number value than the subject but also the verb. In spite of this anomaly the mismatches of type 3 are quite common. The sample, for instance, comprises 153 instances. Nearly all of them are combinations in which the subject is the impersonal neuter pronoun *het* 'it' (76 hits) or one of the demonstrative neuter pronouns, i.e. *dit* 'this' (36 hits) or *dat* 'that' (33 hits). Some examples are given in (48–50).

(48) Het worden spannende maanden.
it become exciting months
'It'll be exciting months.' (dpc-vhs-000759-nl-sen-p.28.s.1)

(49) Dit zijn uiterst verontrustende berichten.
this are extremely worrying messages
'These are extremely worrying messages.'
(dpc-bal-001239-nl-sen-p.60.s.1)

(50) Dat zijn sterke emoties.
that are strong emotions
'Those are strong emotions.' (wr-p-p-c-0000000048.txt-148)

[12] The other example of this kind is *geen domotica* 'no domotics'.

The predicate nominals in (48–50) have singular counterparts that are also compatible with *het, dit* and *dat*.

(51) a. Het wordt/*worden een spannende maand.
it becomes/*become an exciting month
'It'll be an exciting month.'
b. Dit is/*zijn een verontrustend bericht.
this is/*are a worrying message
'This is a worrying message.'
c. Dat is/*zijn een sterke emotie.
that is/*are a strong emotion
'That is a strong emotion.'

Notice that the verb must be singular as well in (51). Given that the verb canonically shows number agreement with the subject, this might be seen as evidence that the subject in (48–51) is the postverbal nominal, rather than the pronoun. Attractive as this option might seem, we do not adopt it for four reasons.

The first is related to the fact that phonologically reduced pronouns, such as *je* 'you' and *het* 'it', can only occur in the preverbal position if they are subjects, as in (52). If they are complements, one has to use their non-reduced counterpart, as shown in (53).

(52) a. Je komt altijd te laat.
you come always too late
'You are always late.'
b. Het is te klein.
it is too small
'it is too small.'

(53) a. Jou/*je had ik nog niet ontmoet.
you had I still not met
'You I hadn't met yet.'
b. Dat/*het wist ik niet.
that/*it knew I not
'That I did not know.'

As a consequence, the fact that the reduced form *het* 'it' takes the preverbal position in (48) and (51a) provides evidence that it is a subject, rather than a preposed predicative complement.

The second reason is related to the linear order in inverted and subordinate clauses. In clauses with VSO order the subject canonically occurs immediately after the finite verb and in subordinate clauses with SOV order it canonically occurs immediately after the complementizer. The fact that the pronouns take that position in the inverted (54–55) and in the subordinate clauses of (56–57), hence, demonstrates that they are the subject.

(54) Meestal zijn het kleine bevingen die geen gevaar opleveren.
mostly are it small shakes that no danger cause
'Most of the time it is small shakes that do not cause any danger.'
(wr-p-e-h-0000000049.p.34.s.2)

(55) In feite waren dit slechts kleine relletjes.
in fact were this only small riots
'These were in fact just small riots.' (wr-p-e-i-0000050381.p.1.s.96)

(56) In het museum ontstond grote paniek omdat men dacht dat het terroristen waren.
in the museum arose great panic because one thought that it terrorists were
'In the museum arose great panic because one thought that it terrorists were.' (ws-u-e-a-0000000232.p.6.s.3)

(57) We weten dat dat leugens zijn.
we know that that lies are
'We know that those are lies.' (ws-u-e-a-0000000210.p.25.s.21)

Further evidence is provided by the fact that the combinations become ill-formed when the plural nominals take the position just after the verb or the complementizer.

(58) a. * Meestal zijn kleine bevingen het die geen gevaar opleveren
* mostly are small shakes it that no danger cause
b. * In feite waren slechts kleine relletjes dit
* in fact were only small riots this
c. * In het museum ontstond grote paniek omdat men dacht dat terroristen het waren.
* in the museum arose great panic because one thought that terrorists it were
d. * We weten dat leugens dat zijn.
* we know that lies that are

The third reason concerns the fact that number mismatches between subject and finite verb are not limited to combinations with *het, dit* and *dat*. They also occur in combinations with a full NP (8 hits), as in (59–60).

(59) De kleding die ze droegen waren vermoedelijk dierenvellen.
the clothing that they wore were probably animal.hides
'The clothing they wore were probably animal hides.'
(wr-p-e-i-0000050381.p.1.s.16)

(60) Een kind kan zien dat het trio van de 'As van het kwaad' toevallig ook de vijanden
A child can see that the trio of the 'Axis of evil' incidentally also the enemies
van Israël zijn.
of Israel are
'Even a child can see that the trio of the 'Axis of evil' are incidentally also the enemies of Israel.'
(wr-p-p-i-0000000172.p.3.s.10)

In these cases it is not only the linear order that favours the assignment of subject status to the singular NPs, but also the thematic structure. In (59), for instance, the plural *dierenvellen* 'animal hides' is part of the VP *waren vermoedelijk dierenvellen* 'were probably animal hides', which expresses a property of the clothing that they wore. A similar remark applies to (60).

The fourth reason is that combinations of a singular subject and a plural verb also occur in other Germanic languages. Pollard and Sag (1994, 70), for instance, provide the following examples from English.

(61) a. The faculty are all agreed on this point.
b. The government are setting new wage standards.
c. If your family are all going to be here next week, then let's have a party.

Notice that these examples resemble the Dutch (60) in that the subject is headed by an inherently collective noun.

In sum, while finite verbs canonically show number agreement with the subject in Dutch, there are some exceptions. These include combinations with the singular neuter pronouns *het* 'it', *dit* 'this' and *dat* 'that', as well as combinations with subjects that are headed by a *singulare tantum*, such as *kleding* 'clothing', or an inherently collective noun, such as *trio*. We, hence, endorse the practice of the annotators and the parser to assign subject status to the singular nominals in such examples as (48–50), (54–55), (56–57) and (59–60).

The predicate nominals in these combinations are mainly bare plurals (96 hits) or definite NPs (50 hits). Indefinite NPs are rare (4 hits). Completing the survey are two NPs with a *wh*-determiner (*om het even welke waarden* 'no matter which values' and *wat voor gerechten* 'what kind of dishes') and one pronoun.

4.4. Singular verb and singular PREDC vs. plural subject

With only seven instances this is the least common type of mismatch in the sample. Some examples are given in (62–63).

(62) Ten tweede was de vraag of de afwijkende loopbanen slechts een voorbijgaand
At second was the question if the deviant careers only a transient
fenomeen is dat typisch is voor de eerste generatie werkende vrouwen...
phenomenon is that typical is for the first generation working women...
'Secondly, the question was whether the deviant careers is only a transient phenomenon that is typical for the first generation of working women...' (dpc-fsz-000551-nl-sen.p.15.s.6)

(63) De Vulcans is een ras van zeer intelligente mensachtigen,
the Vulcans is a race of very intelligent humanoids,
die logica als de basis voor iedere beslissing zien.
who logic as the basis for every decision see
'The Vulcans is a race of very intelligent humanoids, who see logic as the basis of every decision.'
(wr-p-e-i-0000027197.p.3.s.155.2)

Given that the finite verbs in these sentences show number agreement with the singular nominal, it might be argued that the treatment of the plural nominals as subjects is erroneous. For a number of reasons, though, partly similar to those discussed in section 4.3, we consider the subject treatment of the plural nominals to be correct. The first reason is that they occur in positions which are typical for the subject, such as the position just after the complementizer in the subordinate clause of (62). Just like in (58), the reversal of the NPs in that clause yields an ill-formed sequence.

(64) *... of een voorbijgaand fenomeen de afwijkende loopbanen is
*... if a transient phenomenon the deviant careers is

The second reason is that mismatches of this kind also occur in other Germanic languages. Pollard and Sag (1994:85–86), for instance, mention the following example from English.

(65) Eggs is my favorite breakfast.

In sum, we endorse the practice of the annotators and the parser to assign subject status to the plural nominals in (62–63). The predicate nominals in these combinations are full NPs (2 definite and 5 indefinite).

4.5. Summing up

This section has provided an overview of the four types of mismatches that are possible in the relation between subjects, verbs and predicate nominals. The mismatches in clauses with a plural verb (types 1 and 3) are quite common (329 hits) and unexceptional in terms of both well-formedness and acceptability. By contrast, the mismatches in clauses with a singular verb (types 2 and 4) are uncommon (19 hits) and rather marked. (62), for instance, is on the verge of ill-formedness.

5. A unified formal account

This section provides a unified account of the data. It employs the Typed Feature Structure notation of Head-driven Phrase Structure Grammar (HPSG) and takes some inspiration from the latter's treatment of agreement, briefly presented in section 5.1. Another source of inspiration is Peter Lasersohn's treatment of the distinction between distributive and collective interpretations (Lasersohn, 1995). It is used in our account of the first two types of mismatches in section 5.2. The other types, which involve a double mismatch, are dealt with in section 5.3.

5.1. Concord and index agreement

HPSG makes a distinction between morpho-syntactic agreement (also known as concord) and index agreement. The distinction was introduced in chapter 2 of Pollard and Sag (1994) and further developed in Kathol (1999) and Wechsler and Zlatić (2003).[13] The latter defines it in terms of the scheme in (66).

(66) morphology ⇔ CONCORD ⇔ INDEX ⇔ semantics

''We recognize two distinct grammaticalization 'portals', one each via semantics and morphology. These two sources of grammaticalization lead to two distinct bundles of agreement features for a given noun. The morphology-related

[13] The need for a distinction along these lines is also felt in transformational grammar. Sauerland and Elbourne (2002), for instance, employs the NUMBER feature to model morpho-syntactic agreement, and proposes a new feature, called MEREOLOGY, to capture something which resembles index agreement.

agreement bundle will be called CONCORD (which includes case, number and gender) and the semantics-related agreement bundle will be called INDEX (which includes person, number and gender).'' (Wechsler and Zlatić, 2003, 28)

A typical example of concord is the NP-internal agreement in case, number and gender between determiners, prenominal adjectives and nouns in German and Dutch, see Pollard and Sag (1994), Netter (1996) and Kathol (1999) on German, and Van Eynde (2006) on Dutch. A typical example of index agreement is the agreement in person, number and gender between an anaphoric pronoun and its antecedent, see Pollard and Sag (1994) and Sag et al. (2003).

Returning now to the agreement in copular constructions, it has been argued on the basis of examples from the Romance languages that the number agreement between subject and finite verb is an instance of concord, while the number agreement between subject and predicative adjective is an instance of index agreement. Some evidence is provided by the French data in (67).

(67) a. Vous êtes/*es loyal.
you be.2PL/*be.2SG loyal.SG
'You are loyal.'

b. Vous êtes/*es loyaux.
you be.2PL/*be.2SG loyal.PL
'You are loyal.'

The second person pronoun *vous* invariably requires a plural verb, when it is used as a subject, but it is less choosy with respect to the predicative adjective. To model this Kathol (1999:248) assumes that the pronoun has a morpho-syntactic number value (AGR|NUMBER) that is unambiguously plural, while its index is underspecified:[14]

(68)
$$\left[\begin{array}{l} \ldots \mid \text{AGR}\left[\begin{array}{ll}\text{NUMBER} & \textit{plural}\\ \text{GENDER} & \textit{gender}\end{array}\right] \\ \ldots \mid \text{INDEX}\left[\begin{array}{ll}\text{PERSON} & 2\\ \text{NUMBER} & \textit{number}\\ \text{GENDER} & \textit{gender}\end{array}\right] \end{array}\right]$$

This accounts for the data in (67), if one assumes that the agreement between subject and verb is an instance of concord, while the agreement between subject and predicative complement is an instance of index agreement. More specifically, the underspecified INDEX|NUMBER value of the pronoun is resolved to *singular* if *vous* denotes a single addressee, as in (67a), and to *plural* if it denotes a group of addressees, as in (67b).

Taking a cue from Kathol's treatment, we assume that the number agreement between subject and verb is an instance of concord in Dutch. To model it we make use of lexical rules. The formation of the plural forms of the verbs, for instance, is modeled in (69).

(69)
$$\left[\begin{array}{ll}\textit{verb-stem}\\ \text{FORM} & \boxed{1}\end{array}\right] \Rightarrow_{LR} \left[\begin{array}{ll}\textit{word}\\ \text{FORM} & F_{pl}\left(\boxed{1}\right)\\ \text{SUBJECT} & \left\langle \text{NP}\left[\text{AGR} \mid \text{NUMBER}\ \textit{plural}\right]\right\rangle\end{array}\right]$$

[1] stands for a verbal stem, such as *word* 'become', or its past tense counterpart, *werd* 'became'. F_{pl} is a morpho-phonological function that relates it to its plural form. This involves the addition of the suffix *-en*, as in *worden* and *werden*. The resulting form requires its subject to be morpho-syntactically plural. Significantly, the lexical rule does not require the subject to have a plural index.

[14] Kathol's AGR feature corresponds to Wechsler and Zlatic's CONCORD feature. The GENDER value is underspecified for both AGR and INDEX.

There are similar rules for the singular forms. More specifically, there is one for the forms without affix, as in *word* 'become', and one for the forms with the suffix *-t*, as in *wordt* 'becomes'. The former is used for the present tense forms of the first person (*ik word*) and the inverted second person (*word je*), as well as for the past tense forms, as in *werd* 'became'. The latter is used for the present tense forms of the third person (*hij wordt*) and the non-inverted second person (*je wordt*). Both of these rules require the subject to be morpho-syntactically singular.

Notice that (69) and its counterparts for the singular forms only deal with the regular cases. To model the mismatches that were discussed in sections 4.3 and 4.4 we propose an extra set of lexical rules in section 5.3.

Turning to the number agreement between predicate nominals and their target, it is tempting to take a cue from Kathol's treatment of the number agreement between predicative adjectives and their target. More specifically, we could model the agreement in terms of index sharing and allow mismatches for nominals with an underspecified number value in their index, comparable to what is assumed for the French pronoun *vous* 'you' in (68). A problem for this treatment, though, is that predicate nominals—in contrast to predicative adjectives—have their own PERSON and GENDER values. Requiring them to share the index of their target erroneously predicts that combinations such as *you are a genius* (with conflicting PERSON values) and *her father is a woman now* (with conflicting GENDER values) are ill-formed. For more examples of this kind, see Van Eynde (2015:186–191).

In sum, the agreement effect which we observed and quantified in section 3.3 cannot be modeled in terms of the grammaticalized forms of agreement: Concord is not appropriate because of the many mismatches, and index sharing is not appropriate either. At the same time, the agreement effect is undeniable and has to be captured somehow. sections 5.2 and 5.3 present a proposal.

5.2. Partial index agreement

This section deals with the combinations in which the predicate nominal and its target show morpho-syntactic number agreement as well as with the mismatches of type 1 and 2. As a starting point we take the general characterization of the copular verbs in Van Eynde (2015:116), spelled out in (70).[15]

(70)
$$\left[\begin{array}{l} \textit{copular-verb-lexeme} \\ \text{ARGUMENT-STRUCTURE} \left\langle X_i , Y_j \right\rangle \\ \text{CONTENT} \left[\begin{array}{l} \textit{state-of-affairs} \\ \text{NUCLEUS} \left[\begin{array}{l} \textit{attribute-relation} \\ \text{INSTANCE } s \\ \text{THEME } i \\ \text{ATTRIBUTE } j \end{array}\right] \end{array}\right] \end{array}\right]$$

Put in plain words, copular verb lexemes take two syntactic arguments and denote an object of type *state-of-affairs*. They assign the THEME role to the first argument and the ATTRIBUTE role to the second one. The indices of the arguments are associated with PERSON, NUMBER and GENDER features, as in (68). The value of the INSTANCE feature is also an index: It stands for the situation that the verb denotes. We assume that the indices of type *situation* also contain a NUMBER feature and use this to capture the distinction between verbal projections with a distributive interpretation and verbal projections with a collective interpretation. Taking a lead from the treatment of this distinction in Lasersohn (1995), we assign the value *plural* to the projections with a distributive interpretation and the value *singular* to the projections with a non-distributive interpretation.[16]

It may be worth stressing that (70) not only subsumes the copula *zijn* 'be', but also other copular verbs, such as *worden* 'become', *blijven* 'stay' and *lijken* 'seem'. A comprehensive survey of the Dutch verbs with copular uses in the LASSY treebank is provided in Van Eynde et al. (2014). It comprises three dozens of verbs, classified in terms of a number of semantic distinctions, such as stative vs. dynamic and telic vs. atelic.

Turning now to the number agreement between subject and predicate nominal, we first treat the combinations with a singular subject and then those with a plural one.

[15] This is not an exact copy of Figure 46 in Van Eynde (2015:116). On the one hand, (70) omits features which are not relevant in the present context. On the other hand, it adds the INSTANCE feature, which is not explicitly mentioned in the original, because it is inherited from a supertype of *attribute-relation*.

[16] Collective interpretations are a proper subset of the non-distributive interpretations.

5.2.1. Combinations with a singular subject

To model the agreement in combinations with a singular subject we employ the phrasal constraint in (71).

(71)
$$\left[\begin{array}{l} \textit{head-complements-phrase} \\ \text{DAUGHTERS} \left\langle \mathbf{H} \text{ , } \text{NP}_j \right\rangle \\ \text{CATEGORY}|\,\text{SUBJECT} \left\langle \text{NP}_i \left[\text{AGR}|\,\text{NUMBER} \quad \textit{singular}\right]\right\rangle \\ \text{CONTENT}|\,\text{NUCLEUS} \left[\begin{array}{l} \textit{attribute-relation} \\ \text{INSTANCE} \quad s \\ \text{THEME} \quad i \\ \text{ATTRIBUTE} \quad j \end{array}\right] \end{array}\right]$$

$$\Rightarrow \left[\text{CONTENT}|\,\text{NUCLEUS} \left[\begin{array}{ll} \text{INSTANCE}|\,\text{NUMBER} & \boxed{1}\ \textit{singular} \\ \text{THEME}|\,\text{NUMBER} & \boxed{1} \\ \text{ATTRIBUTE}|\,\text{NUMBER} & \boxed{1} \end{array}\right]\right]$$

(71) is an implicational constraint. It subsumes the signs whose properties are spelled out in the left hand side of the arrow and requires such signs to also have the properties which are spelled out in the right hand side. In this case the subsumed signs are phrases

1. which contain a head daughter (**H**) and its nominal complement,
2. which select a singular subject, and
3. whose complement daughter is assigned the ATTRIBUTE role.

In practice, these are VPs which contain a singular finite verb and a predicate nominal. Of such signs, (71) requires that the number values in the indices of the subject, the predicate nominal and the VP are all singular.

Assuming that predicate nominals with a singular index are also morpho-syntactically singular, the constraint in (71) accounts for the combinations in which the subject and the predicate nominal are both singular, which is in fact the large majority of the combinations in the sample. In addition, the constraint also accounts for the mismatches of type 2. As observed in section 4.2, there are twelve instances of that type in the sample: nine with a numeral, two with a *plurale tantum* and one with a topicalized predicate nominal.

The combination with the plural topicalized predicate nominal in (47) is not subsumed by the left hand side of (71), since it is not an instance of the *head-complement-phrase* type. In terms of the HPSG hierarchy of phrase types, the combination of a topicalized constituent with a gapped clause is subsumed by the *head-filler-phrase* type.

The combinations with a predicate nominal that is headed by a *plurale tantum*, such as *activa* 'activa' or *domotica* 'domotics', are subsumed by the left hand side of (71) and comply with its right hand side, if one makes the reasonable assumption that *pluralia tantum* have an underspecified number value in their index. In that respect, they resemble the French pronoun *vous* in (68).

The combinations with a predicate nominal that contains a numeral, as in (45), repeated in (72), are also subsumed by the left hand side of (71).

(72) Bij een vrouw is de grens veertien glazen.
with a woman is the limit fourteen glasses
'For a woman the limit is fourteen glasses.'

At first blush, they do not seem to comply with the right hand side. A closer look, though, reveals that they have a peculiar interpretation: (72), for instance, says something about the number of glasses rather than about the glasses as such. This is a property which it shares with (73), where the finite verb has a singular form.

(73) Veertien glazen is ruim voldoende.
fourteen glasses is amply sufficient
'Fourteen glasses is amply sufficient.'

A plausible way to account for this is to assume that nominals which contain a numeral are headed by that numeral if the salient information concerns the quantity rather than the substance of whatever it is that the nominal denotes. As a

consequence, if we assume that the Dutch numerals are singular count nouns, as argued in Van Eynde (2006) on the basis of observations about their morphology and their distribution, then the predicate nominal in (72) is a bare singular, rather than a bare plural, and in that case it is compatible with the constraint in (71).[17] By the same token, this accounts for the compatibility with the singular finite verb in (73).

5.2.2. Combinations with a plural subject

The combinations with a plural subject are more diverse. This is not surprising, for while the matches vastly outnumber the mismatches in the combinations with a singular subject (2801 matches vs. 12 mismatches), they account for less than half of the combinations with a plural subject (158 matches vs. 176 mismatches). Besides, while the combinations with a singular subject all have a non-distributive interpretation, the combinations with a plural subject can be either distributive or non-distributive. To account for the role of this distinction we add it to the left hand side of the constraint in (74).

(74)
$$\left[\begin{array}{l} \textit{head-complements-phrase} \\ \text{DAUGHTERS} \ \langle \mathbf{H} \ , \ \text{NP}_j \rangle \\ \text{CATEGORY} \mid \text{SUBJECT} \ \langle \text{NP}_i \ [\text{AGR} \mid \text{NUMBER} \ \textit{plural}] \rangle \\ \text{CONTENT} \mid \text{NUCLEUS} \left[\begin{array}{l} \textit{attribute-relation} \\ \text{INSTANCE} \mid \text{NUMBER} \ \boxed{1} \ \textit{plural} \\ \text{THEME} \ \ i \\ \text{ATTRIBUTE} \ \ j \end{array}\right] \end{array}\right]$$
$$\Rightarrow \left[\text{CONTENT} \mid \text{NUCLEUS} \left[\begin{array}{ll} \text{THEME} \mid \text{NUMBER} & \boxed{1} \\ \text{ATTRIBUTE} \mid \text{NUMBER} & \boxed{1} \end{array}\right]\right]$$

The left hand side subsumes signs of type *head-complements-phrase*

1. which contain a head daughter (**H**) and its nominal complement,
2. which select a plural subject,
3. whose complement daughter is assigned the ATTRIBUTE role, and
4. which have a distributive interpretation.

The right hand side requires the subject and the predicate nominal in such phrases to have a plural index.

Assuming that predicate nominals with a plural index are morpho-syntactically plural, the constraint in (74) models the combinations in which the subject and the predicate nominal are both morpho-syntactically plural, as in the first sentence of (24), repeated in (75).

(75) Dat betekent niet dat de initiatiefnemers nu ineens managers zijn.
that means not that the initiators now suddenly managers are.
Ze zijn en blijven vooral boer.
they are and remain chiefly farmer
'That does not mean that the initiators are now all of a sudden managers. They are and remain in the first place farmers.'

(74) also models the type 1 mismatches with a distributive interpretation, as in the second sentence of (75), if one makes the reasonable assumption that singular bare predicate nominals, such as *boer* 'farmer', have an underspecified index. This assumption, in fact, accounts for the observation, made in section 4.1, that the singular and plural forms of the predicate nominals are interchangeable without change of meaning in combinations with a distributive interpretation. Because of the underspecification the NUMBER value in the index is resolvable to *plural*, matching the NUMBER value in the index of the subject. The resulting discrepancy between a singular morpho-syntactic number and a plural index is possible

[17] In this respect, we propose another analysis than the one of the LASSY annotators. We do not count the latter as erroneous, though, since the annotation makes good sense for the canonical uses of nominals with a numeral.

for the bare singular predicate nominals as well as for the saturated predicative NPs which denote a role rather than an entity or a group, see section 4.1.

The constraint in (74) does not cover the type 1 mismatches with a collective interpretation. This is deliberate, since such combinations are exempt from number agreement. Moreover, just like (71), the constraint does not subsume the combinations with a topicalized predicate nominal. This chimes well with the fact that such combinations have a collective interpretation.

5.2.3. Summing up

The resulting treatment is not based on either concord or index sharing, but on partial index agreement. It is partial in an obvious sense, since it only concerns the NUMBER values of the indices, but it is also partial in a more subtle sense, since the constraint on the combinations with a plural subject is explicitly limited to those with a distributive interpretation. The phrasal constraints which model the agreement are clearly disjoint, since they put mutually exclusive constraints on the morpho-syntactic number of the subject, but they do not jointly subsume all the possible combinations. Combinations with a topicalized predicate nominal, for instance, are not subsumed. Neither are combinations with a collective interpretation. Moreover, the constraints require the subject to have the same value for AGR|NUMBER and INDEX|NUMBER. That this need not always be the case will be shown in section 5.3.

5.3. The double mismatches

The mismatches of type 3 not only show number discord between the predicate nominal and the subject, but also between the verb and the subject. This means that we cannot use the lexical rule for plural verbs in (69) to model its use in (48), repeated in (76).

(76) Het worden spannende maanden.
it become exciting months
'It'll be exciting months.'

To accommodate such combinations we add a second lexical rule for the plural forms of copular verbs.

(77)
$$\begin{bmatrix} \textit{copular-verb-stem} \\ \text{FORM } \boxed{1} \\ \text{ARG-ST } \left\langle \boxed{2}\text{NP}, \boxed{3}\text{NP} \right\rangle \end{bmatrix} \Rightarrow_{LR} \begin{bmatrix} \textit{word} \\ \text{FORM } F_{pl}\left(\boxed{1}\right) \\ \text{ARG-ST } \left\langle \boxed{2}\text{NP}\begin{bmatrix} \text{AGR} \mid \text{NUMBER} & \textit{singular} \\ \text{INDEX} \mid \text{NUMBER} & \boxed{4}\textit{plural} \end{bmatrix}, \boxed{3}\text{NP}\left[\text{INDEX} \mid \text{NUMBER } \boxed{4}\right] \right\rangle \end{bmatrix}$$

The range of application of this rule is much smaller than that of (69), since it only applies to copular verbs. Moreover, it only applies to those uses of the copular verbs in which their second argument is an NP. This excludes among others combinations with predicative adjectives and PPs, and thus accounts for the fact that (76) is well-formed, while *het zijn spannend* 'it are exciting' is not. For the signs which are subsumed by the left hand side of (77), the rule stipulates that their plural forms require a singular first argument with a plural index and a second argument with a plural index. The discrepancy between morpho-syntactic number and index number in the first argument is possible for singular (pro)nouns whose index is underspecified for number. They include the impersonal pronoun *het* 'it', the neuter demonstrative pronouns *dit* 'this' and *dat* 'that', *singularia tantum*, such as *kleding* 'clothing', and inherently collective nouns, such as *trio* 'trio'.

It may be worth stressing that this underspecification differs from the one that we observed for the demonstrative *die* 'that/those' in section 2. While *die* has an underspecified AGR|NUMBER value, so that it is compatible with singular and plural VPs alike, as shown in (5), *dat* and its companions have a specific AGR|NUMBER value (*singular*), but an underspecified INDEX|NUMBER value. Their compatibility with plural verbs is, hence, limited to copular verbs with a plural predicate nominal.

Another member of the class of singular pronouns with an underspecified index is the interrogative *wat* 'what'. Combinations with this pronoun are not in the sample, since interrogative pronouns are canonically preposed and, hence, assigned the dependency label WHD, rather than SU or PREDC, but they are worth a closer look, since their combination with *het, dit, dat* shows a remarkable property, observed in Perridon (2014:69) and brought to our attention by an anonymous reviewer.

(78) a. Wat is dat/dit/het ?
what is that/this/it ?
'What's that/this/it?'
b. Wat zijn dat/dit/het ?
what are that/this/it ?
'What are those/these?'

(78a) asks something about one instance of whatever it is that *dat/dit/het* denotes, while (78b) asks something about multiple instances. Both combinations are well-formed, and, interestingly, both are licensed by our treatment. The singular verb in (78a) is modeled by one of the lexical rules for singular verb forms and, hence, requires a subject that is morpho-syntactically singular. If *wat* 'what' is the subject, the constraint on the VP in (71) requires the indices of the pronouns to be singular too.[18] The plural verb in (78b) is modeled by the lexical rule in (77). It requires both arguments to have a plural index and the first argument to be morpho-syntactically singular. This licenses both the case in which *wat* 'what' is the subject and the case in which *wat* 'what' is the preposed predicate nominal.

For the sake of completeness we also return to the mismatches of type 4. A relevant example is (63), repeated in (79).

(79) De Vulcans is een ras van zeer intelligente mensachtigen,
the Vulcans is a race of very intelligent humanoids,
die logica als de basis voor iedere beslissing zien.
who logic as the basis for every decision see
'The Vulcans is a race of very intelligent humanoids, who see logic as the basis of every decision.'

This combination is licensed by a lexical rule that is the mirror image of (77): Its output allows the singular forms of copular verbs to select arguments with a singular index even though their first argument is morpho-syntactically plural. Since the latter deviates from the general tendency for plural nouns to have a plural index, it is not surprising that only certain nominals allow it and that mismatches of this type are few and far between (7 hits).

5.4. Wrapping up

To model the agreement effect in copular constructions with a predicate nominal we have built on the distinction between morpho-syntactic agreement and index agreement, familiar from HPSG, as well as on the treatment of the distinction between distributive and collective interpretations in Lasersohn (1995). The resulting treatment consists of

- lexical rules for number concord between subject and verb, as in (69)
- phrasal constraints on partial index agreement, as in (71) and (74)
- lexical rules for number discord, as in (77)

The lexical rules for number concord are independently needed for the other verbs. The phrasal constraints model the combinations with matching number values as well as the mismatches of type 2 and those of type 1 that have a distributive interpretation. The lexical rules for number discord model the mismatches of type 3 and 4.

6. Conclusion

The research on which this article reports has a dual aim. The methodological aim is to demonstrate how treebanks can be used to guide the formulation of generalizations. For that purpose we made use of tools and resources that have recently become available in the framework of the STEVIN program and the CLARIN infrastructure.

The theoretical aim is to enhance our understanding of the number agreement phenomenon in copular constructions. Building on work in Head-driven Phrase Structure Grammar and Formal Semantics, we have developed a unified account of the phenomenon, which is made explicit in terms of a small number of lexical rules and phrasal constraints.

Acknowledgements

The data-oriented part of the paper has been presented at a number of workshops and conferences, including the 24th CLIN Conference in Leiden (2014) and the LOT Summer School in Nijmegen (2014). We wish to thank the

[18] If *wat* 'what' is the (preposed) predicate nominal, the sentence is not subsumed by (71).

audiences at these occasions and the anonymous reviewers of this paper for their comments and suggestions for improvement.

References

Augustinus, L., Vandeghinste, V., Van Eynde, F., 2012. Example-based treebank querying. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), pp. 3161–3167.

Fillmore, C., Lee-Goldman, R.L., Rhomieux, R., 2012. The FrameNet construction. In: Boas, H.C., Sag, I.A. (Eds.), Sign-based Construction Grammar. CSLI Publications, Stanford University, pp. 309–372.

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

Jongejan, B., Olsen, S., Fersoe, H., 2011. Validation Report Lassy Corpora Linguistic Validation, Technical report. Center for Sprogteknologi, University of Copenhagen.

Kathol, A., 1999. Agreement and the syntax-morphology interface in HPSG. In: Levine, R.D., Greene, G.M. (Eds.), Studies in Contemporary Phrase Structure Grammar. Cambridge University Press, Cambridge, pp. 223–274.

Lasersohn, P., 1995. Plurality, Conjunction and Events. Kluwer Academic Publishers, Dordrecht.

Netter, K., 1996. Functional Categories in an HPSG for German (PhD thesis). Universität des Saarlandes, Saarbrücken.

Perridon, H., 2014. Enige opmerkingen over de vorm van persoonlijke voornaamwoorden in koppelzinnen. In: Van de Velde, F., Smessaert, H., Van Eynde, F., Verbrugge, S. (Eds.), Patroon en argument. Universitaire Pers, Leuven, pp. 63–74.

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

Sag, I.A., Wasow, T., Bender, E., 2003. Syntactic Theory. A Formal Introduction, 2nd ed. CSLI Publications, Stanford University.

Sauerland, Uli, Elbourne, Paul, 2002. Total reconstruction, PF movement and derivational order. Linguist. Inq. 33, 283–319MIT Press.

Spyns, P., Odijk, J. (Eds.), 2013. Essential Speech and Language Technology for Dutch. Springer, Berlin.

Van Eynde, F., 2003. Part of speech tagging en lemmatisering van het Corpus Gesproken Nederlands, Leuven.

Van Eynde, F., 2006. NP-internal agreement and the structure of the noun phrase. J. Linguist. 42, 139–186.

Van Eynde, F., 2015. Predicative Constructions. From the Fregean to a Montagovian Treatment. CSLI Publications, Stanford University.

Van Eynde, F., Augustinus, L., Schuurman, I., Vandeghinste, V., 2014. Het verrassende resultaat van een copulativiteitspeiling. In: Van de Velde, F., Smessaert, H., Van Eynde, F., Verbrugge, S. (Eds.), Patroon en argument. Universitaire Pers Leuven, Leuven, pp. 47–62.

Van Noord, Gertjan, 2006. At Last Parsing is Now Operational. In: Mertens, P., Fairon, C., Dister, A., Watrin, P. (Eds.), TALN 2006. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles. pp. 20–42.

Van Noord, Gertjan, Bouma, Gosse, Van Eynde, Frank, De Kok, Daniel, Van der Linde, Jelmer, Schuurman, Ineke, Kim Sang, Erik Tjong, Vandeghinste, Vincent, 2013. Large scale syntactic annotation of written Dutch: Lassy. In: Spyns, Peter, Odijk, Jan (Eds.), Essential Speech and Language Technology for Dutch. Springer, Berlin, pp. 147–164.

Wechsler, S., Zlatić, L., 2003. The Many Faces of Agreement. CSLI Publications, Stanford University.