Number agreement in copular constructions
A treebank-based investigation

Frank Van Eynde
Center for Computational Linguistics
KU Leuven

LOT Summer School 2014
CLARIN for Linguists

THE PHENOMENON

(1) a. Zijn broer is een schurk.
b. Zijn broers zijn schurken.

(2) a. * Zijn broers zijn een schurk.
b. * Zijn broer is schurken.

(3) a. Ik ben beste maatjes met de president van Finland.
b. Zijn broers zijn een gevaar voor de maatschappij.

Questions:

- how common are the mismatches?
- under which circumstances do the mismatches occur?

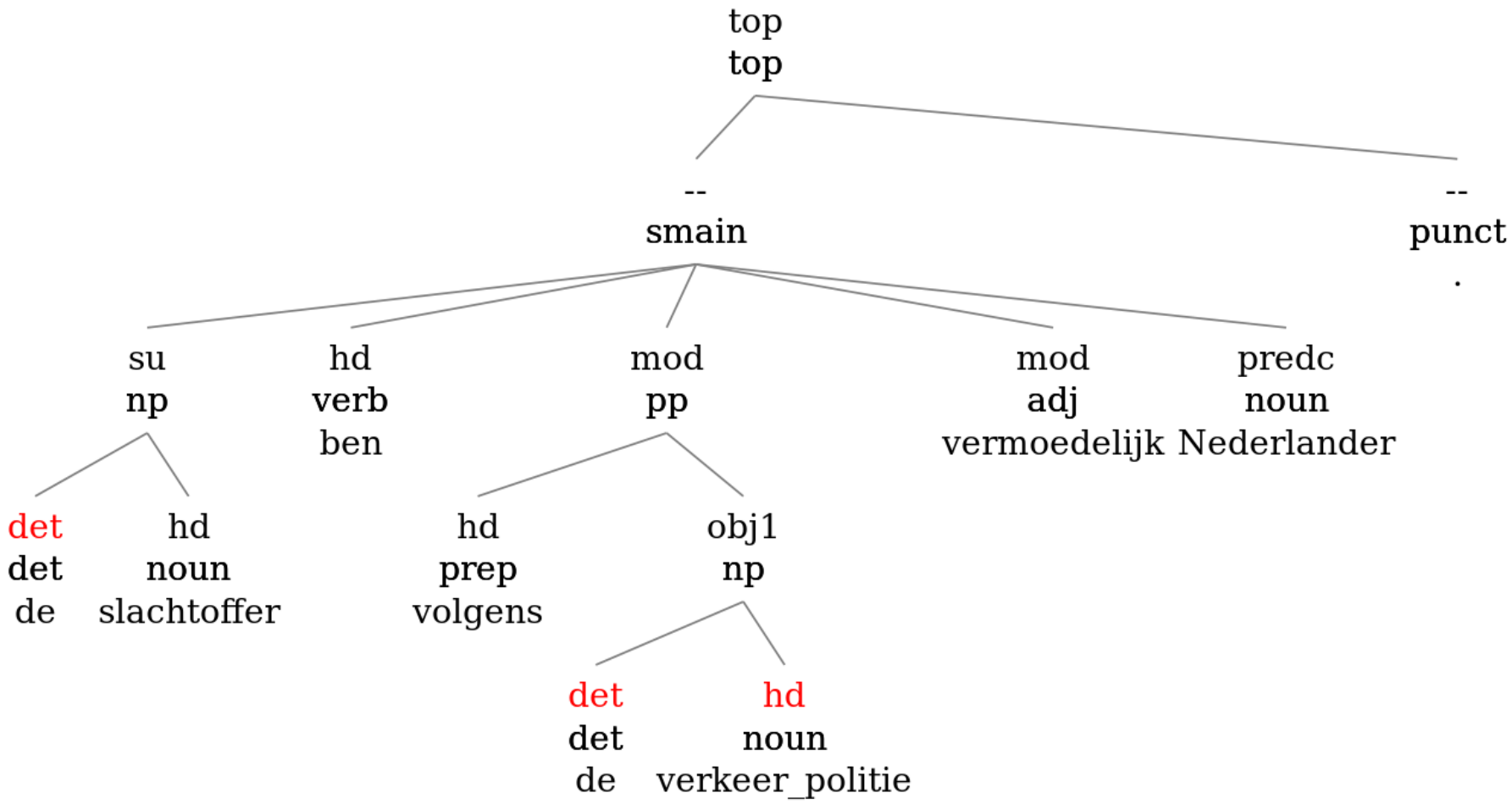
COMPOSITION AND SIZE OF LASSY SMALL

Treebank	Contents	# sent	# 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, text for the visually impaired	14,032	184,611
dpc	Dutch Parallel Corpus	11,716	193,029
wikipedia	Dutch Wikipedia pages	7,341	83,360
Total		65,200	975,055

(4) De slachtoffers zijn volgens de verkeerspolitie vermoedelijk Nederlanders. [ws-u-e-a-0000000205.p.18.s.2]

DEPENDENCY TREES

De slachtoffers zijn volgens de verkeerspolitie vermoedelijk Nederlanders .



ADDITION OF DETAILED POS TAGS

Number values are added to nouns and pronouns.

	Noun	Pronoun	Sum	%
Singular	188,297	25,900	214,197	71.30
Plural	58,458	8,265	66,723	22.21
Underspecified	0	19,486	19,486	6.49
Sum	246,755	53,651	300,406	

The underspecified value is assigned to some of the personal, demonstrative and reflexive pronouns (*u, die, zich, ...*).

(5) Die komt niet.

(6) Die komen niet.

QUERYING THE TREEBANK

(7) //node[node[**@rel="hd" and @pt="ww"**] and
node[**@rel="su"**] and
node[**@rel="predc"**]] (17903 hits)

Identifying the combinations with a number value for both subject and predicative complement

(8) //node[node[@rel="hd" and @pt="ww"] and
node[@rel="su" and **@getal**] and
node[@rel="predc" and **@getal**]] (164 hits)

(9) //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**]]]
(1527 hits)

CHECKING AGREEMENT

(10) //node[node[@rel="hd" and @pt="ww"] and
node[@rel="su" and **@getal="ev"**] and
node[@rel="predc" and **@getal="ev"**]] (130 hits)

(11) //node[node[@rel="hd" and @pt="ww"] and
node[@rel="su" and **@getal="mv"**] and
node[@rel="predc" and **@getal="mv"**]] (7 hits)

SU–PREDC	sg-sg	sg-und	sg-pl	pl-sg	pl-pl	und-x	Sum
X–X	**130**	2	8	12	**7**	5	164
XP–X	79	0	11	19	18	2	129
X–XP	1640	11	142	53	46	23	1915
XP–XP	1272	4	22	137	90	2	1527
Sum	3121	17	183	221	161	32	3735
%	83.56	0.45	4.90	5.92	4.31	0.86	

ELIMINATING FALSE OR IRRELEVANT HITS

1. Object-oriented PREDCs of verbs like *noemen, vinden* and *maken*

(12) De Limburgers noemen het vierjaarlijkse Wereld Muziek Concours in Kerkrade *het Wimbledon van de blaasmuziek*.

2. Disfluencies

(13) Ook voor de bekendmaking van het beleid aan de mensen en de politieke discussies in de samenleving zijn ze *groot belang*.

ELIMINATING FALSE OR IRRELEVANT HITS - CONT.

3. Annotation errors

3.1. Type 1

(14) Want ongeacht rang of stand, schutters zijn *gelijk*.

(15) De ambtenaar is *lui*.

3.2. Type 2

(16) Het is *koren op de molen* van de terroristen.

(17) De hoofdstad van Wallonië is *Namen*.

ELIMINATING FALSE OR IRRELEVANT HITS - CONT.

4. Friction between annotation guidelines and the purpose of the investigation

4.1. Underspecified NUMBER value for non-pronominal NPs

(18) Vooral in de katoensector is *de VS* de grootste en meest schadelijke subsidieverstrekker.

(19) Vooral in de katoensector is/zijn *de VS* niet meer competitief.

4.2. Autoreferential NPs

(20) Het thema dit jaar is *"Steden"*.

(21) "Steden" lijkt/*lijken me wel een geschikte titel voor dit boek.

ELIMINATING FALSE OR IRRELEVANT HITS - CONT.

4.3. Identification of the head in NPs.

(22) Bij een vrouw is de grens *veertien glazen*.

(23) Veertien glazen zijn tijdens de verhuis gebroken.

(24) Veertien glazen is ruim voldoende.

(25) Veertien is/*zijn mijn geluksgetal.

RESULT

	sg-sg	sg-und	sg-pl	pl-sg	pl-pl	und-x	Sum
	3121	17	183	221	161	32	3735
Obj-Or PredC	-26	0	-3	-21	-2	-2	-54
Disfluency			0	-1			-1
Ann. Error 1			-7	-12			-19
Ann. Error 2	3		3	-6			0
Friction	15		-19	-1		5	0
Result	3113	17	157	180	159	35	3661
%	85.03	0.46	4.29	4.92	4.73	0.96	

How common are the mismatches?

- The mismatches account for 9.21 % of the 3661 relevant occurrences.
- In clauses with a singular subject (3287) the predicate nominal is plural in 4.78 % of the relevant cases. This is well below the average frequency of plurals (22.21 %).
- In clauses with a plural subject (339) the predicate nominal is singular in 53.10 % of the relevant cases. This is below the average frequency of singulars (71.30 %), but it is surprisingly high.

FOUR TYPES OF MISMATCHES

Second question: under which circumstances do the mismatches occur?

	sg-pl	pl-sg	Sum	%
plural verb	155	174	329	97.63
singular verb	2	6	8	2.37
Sum	157	180	354	

(26) Type 1: Het worden spannende maanden. (155)

(27) Type 2: Goud blijft de belangrijkste financiële activa van bijna alle centrale banken. (2)

(28) Type 3: De verkiezingen waren een gemiste kans. (174)

(29) Type 4: ... de vraag of de afwijkende loopbanen slechts een voorbijgaand fenomeen is (6)

TYPE 1: SINGULAR SUBJECT VS. PLURAL VERB AND PLURAL PREDC

The subject is the neuter impersonal pronoun (*het* or *'t*) or a neuter demonstrative pronoun (*dit* or *dat*) (151 hits)

(30) *Het* worden spannende maanden.

(31) *Dit* zijn uiterst verontrustende berichten.

(32) Zijn vrouwen en zijn verleden: *dat* zijn de belangrijkste inspiratiebronnen.

(33) *Het* wordt/*worden spannend.

(34) *Het* wordt/*worden een spannende maand.

The subject is headed by a mass noun or a collective noun (4 hits)

(35) *De kleding die ze droegen* waren vermoedelijk dierenvellen.

(36) *EVISTA* zijn gele ovaalvormige tabletten.

(37) ... dat *het trio van de 'As van het kwaad'* toevallig ook de vijanden van Israël zijn,

(38) *De kleding die ze droegen* was/*waren versleten.

(39) *EVISTA* is/*zijn duur.

(40) ... dat *het trio van de 'As van het kwaad'* niet meer gevaarlijk is/*zijn,

TYPE 2: SINGULAR SUBJECT AND SINGULAR VERB VS. PLURAL PREDC

The predicate nominal lacks the singular form

(41) Goud blijft *de belangrijkste financiële activa van bijna alle centrale banken*.

(42) Anders is het *geen domotica*.

(43) Goudreserves blijven *de belangrijkste financiële activa van bijna alle centrale banken*.

(44) Stofzuigers waren *de meest geavanceerde domotica van die tijd*.

TYPE 3: PLURAL SUBJECT AND PLURAL VERB VS. SINGULAR PREDC

Two subtypes: collective vs. distributive

(45) De verkiezingen waren *een gemiste kans*.

(46) Beide aftredende bestuurders blijven wel *aandeelhouder*.

SUBTYPE 3.1: COLLECTIVE

Inherently collective: *groep, groepering, verzameling, ...*

(47) Is het omdat wij *een volk van bierdrinkers* zijn dat Belgische vorsers zich zo frequent - en met succes - over leverziekten buigen?

(48) .. dat zij *de beste brassband* zijn *van Nederland*.

Prdicate nominals with a unique referent

(49) ... omdat in de plannen de roltrappen *de enige vluchtweg uit de ondergrondse* zijn.

(50) De Leien (Frankrijklei, Italiëlei, Amerikalei, Britselei) zijn *de belangrijkste verkeersader binnen Antwerpen*.

(51) De kernen Heukelom en Montenaken werden *de gemeente Vroenhoven*

Topicalized predicate nominals

(52) *Grote winnaar bij de verkiezingen van 18 mei 2003* waren de socialisten.

(53) *Een heel specifiek Brussels fenomeen* zijn de 22 gemeenschapscentra, die de lokale draaischijf vormen van het Vlaamse sociale en culturele leven.

(54) *Het hoogtepunt in haar sportcarrière* waren de vier gouden medailles die ze won bij de Olympische Spelen van 1948.

Mutually confirming properties

(55) Onder deze laatsten waren de Grieken *de grootste groep*.

(56) *Het grootste probleem tijdens de wedstrijden* zijn de spreekkoren.

SUBTYPE 3.2: DISTRIBUTIVE

Quantified subjects

(57) Beide aftredende bestuurders blijven wel *aandeelhouder*.

(58) Niet alle hoofdcommissarissen zijn werkelijk *hoofdcommissaris*.

(59) Hiervan zijn tevens zes Belgische Europarlementariërs *lid*.

(60) Volgens sommige bronnen werden minstens 156 mensen hiervan *het slachtoffer*.

Other quantifiers

(61) De Arabische staten die onder Brits bewind hadden gestaan werden veelal *een monarchie*.

(62) We zijn allemaal *het slachtoffer van de platonische manier van denken in tweedelingen*.

(63) Zij had ... gemerkt dat vrouwen vaak *het eerste slachtoffer* waren van de politieke instabiliteit en het aanhoudende geweld.

Mediated by anaphora

(64) Zo zullen steeds minder jongemannen zichzelf in een volgende generatie ervan kunnen overtuigen dat ze *"een goede moslim"* zijn als ze onschuldige medemensen afmaken.

The wider context

(65) Dat betekent niet dat de initiatiefnemers nu ineens managers zijn. Ze zijn en blijven vooral *boer*.

(66) Overigens zullen de drempels niet gelden voor werknemers uit Malta en Cyprus. Die eilanden worden per 1 mei óók *EU-lidstaat*.

TYPE 4: PLURAL SUBJECT VS. SINGULAR VERB AND SINGULAR PREDC

(67) *De Vulcans* is een ras van zeer intelligente mensachtigen, die logica als de basis voor iedere beslissing zien.

(68) ... de vraag of *de afwijkende loopbanen* slechts een voorbijgaand fenomeen is

They have a collective interpretation.

Conclusions - 1

- Predicate nominals canonically show number agreement with the subject, but mismatches are not excluded.
- The availability of treebanks provides us with an opportunity to investigate the phenomenon in actual language use.
- Treebank search by means of XPath queries yields a wealth of relevant data, both quantitative and qualitative.
- False and irrelevant hits must be set aside: object-oriented predicate nominals, disfluencies, annotation errors, friction between annotation guidelines and aim of the investigation.

CONCLUSIONS - 2

- In clauses with a singular subject the predicate nominal is plural in 4.78 % of the relevant cases.
- The mismatches are limited to specific combinations: impersonal *het*, demonstrative *dit* and *dat*, singularia tantum, pluralia tantum, collective nouns.
- In clauses with a plural subject the predicate nominal is singular in 53.10 % of the relevant cases.
- The mismatches are more diverse and relate to the distinction between collective and distributive interpretations.
- The analysis of the mismatches provides useful data for theoretical and descriptive linguistics.

References

- Frank Van Eynde, On the agreement between predicative complements and their target. In: Stefan Müller (ed.), *Proceedings of the 19th International Conference on Head-driven Phrase Structure Grammar.* CSLI Publications, Stanford, 2012, pp. 349-367. (http://csli-publications.stanford.edu/HPSG/2012)
- NN, Number agreement in copular constructions. A treebank-based investigation. 35 pages. Submitted to *Lingua*.

Thank you !