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.