Asymmetries in Extraction From Nominal Copular Sentences: a Challenging Case Study for NLP Tools

Paolo Lorusso, Matteo Greco, Cristiano Chesi, Andrea Moro
NEtS at Scuola Universitaria Superiore IUSS.
P.zza Vittoria 15, I-27100 Pavia (Italy)
`{paolo.lorusso, matteo.greco,`
`andrea.moro, cristiano.chesi}@iusspavia.it`

Abstract

In this paper we discuss two types of nominal copular sentences (Canonical and Inverse, Moro 1997) and we demonstrate how the peculiarities of these two configurations are hardly considered by standard NLP tools that are currently publicly available. Here we show that example-based MT tools (e.g. Google Translate) as well as other NLP tools (UDpipe, LinguA, Stanford Parser, and Google Cloud AI API) fail in capturing the critical distinctions between the two structures in the end producing both wrong analyses and, possibly as a consequence of a non-coherent (or missing) structural analysis, incorrect translations in the case of MT tools. To support the proposed analysis, we present also an empirical study showing that native speakers are indeed sensitive to the critical distinctions. This poses a sharp challenge for NLP tools that aim at being cognitively plausible or at least descriptively adequate (Chowdhury & Zamparelli 2018).

1. Introduction

The main hypothesis of this paper is that sentence comprehension cannot be achieved independently from a coherent structural analysis. To support this claim, we first present a precise structural analysis that is critical for recovering the relevant dependencies within specific constructions, then we will show that the crucial structural properties captured by the theoretical framework are in fact correctly perceived by native speakers, but not

revealed by some widely used Natural Language Processing (NLP) tools. This leads to poor performance in tasks like Machine Translation (MT).

This argument seems to us especially relevant in those structural configurations in which a non-local dependency must be established: in parsing, for instance, interpreting correctly a *wh*-dependency requires that the *dependent* (the *wh*-phrase) and the *dependee* (the head selecting the *wh*- phrase as its argument/modifier) are identified, and the nature of the dependence disambiguated (e.g. argument vs. modifier). In (1) we exemplify the special case of a non-local dependency between a *wh*- PP and a DP it depends on (a co-indexed underscore signals the possible extraction sites, hence the dependent constituent; the diacritic "*" prefixes, as usual, illegal sites):

(1) [Di quale segnale]$_i$ [i telescopi *_ $_i$] hanno
 Of which signal   the telescopes   have
  scoperto *_$_i$ [un'interferenza _$_i$]?
  discovered   an interference?
 '[which signal]$_i$ did the telescopes discover an interference of _$_i$?'

The second DP *un'interferenza* (an interference) (the internal argument) is the dependee of the *wh*-phrase and neither the subject DP nor the predicate can host this *wh*- dependency instead.

According to Google Translate (as of 12th July 2019), this second option seems indeed a viable one:

(2) What signal did the telescopes find an interference?

The translation is ill formed being the internal argument of *find* filled both by the *wh*- phrase and

the DP *an interference* (which cannot take a *wh-*DP as its own argument due to the absence of a relevant preposition).

In this work we focus on a similar non-local dependency involving two kinds of copular sentences: Inverse (3.a) and Canonical (3.b). Using these constructions, we will test the availability of *wh-* PP sub-extraction from both the first and the second DP as exemplified in (4).

(3) a. le foto del muro **sono** la causa della rivolta
the pictures of the wall **are** the cause of the riot
b. la causa della rivolta **sono** le foto del muro
the cause of the riot   are   the pictures of-the wall
'the cause of the riot **is** the pictures of the wall'

(4) a. [Di quale rivolta]$_i$ le foto del muro **sono**
of which riot   the pictures of_the wall **are**
la causa _$_i$?
the cause
b. [Di quale muro]$_i$ le foto _ $_i$ **sono**
of which riot   the pictures of the wall **are**
la causa della rivolta?
the cause of_the riot

In the first part of this paper (§2), we will briefly present an analysis for these constructions, then we will demonstrate that native speakers are selectively sensitive both to the copular structural configuration (Canonical vs. Inverse) and to the extraction site (subject vs. predicate) (§3). In §4 we will test the insensibility of some freely available NLP tools (Google Translate, the Natural Language service of Google Cloud AI API, UDpipe, Stanford Parser and LinguA) to the syntactic oppositions previously discussed.
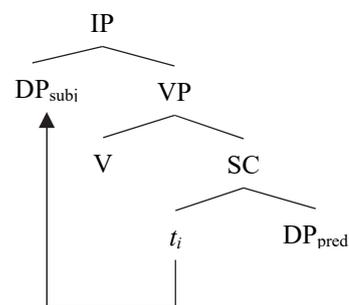
2. The structure of nominal copular sentences

Copular sentences are those sentences whose main verb is *to be* (the copula) and its equivalents across languages. A subset of copular sentences is the one involving two DPs, linearly ordered as DP V DP. Those are dubbed *nominal copular sentences*. In this configuration, a nominal phrase realizes the predicate of the sentence ("the cause…" in (3)) while the other is the subject of the predicate ("the pictures…" in (3)). According to Moro (1997), nominal copular sentences can be distinguished in two subtypes: *Canonical copular* sentences (3.a) – in which the order is subject-copula-predicative expression – and *Inverse copular* sentences (3b) – in which the order is inverted, i.e. predicative expression-copula-subject.
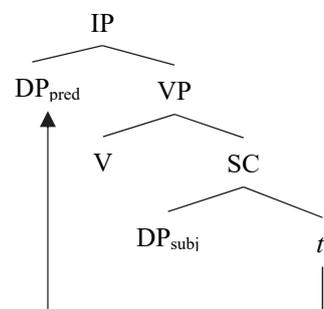
Moro (1991, 1997, 2006) showed that these two types of copular constructions can be distinguished on the basis of different diagnostics like agreement on the verb, grammaticality for the extraction of DPs (Wh- or clitic) and pronominal binding.

Traditionally, copular sentences are analyzed as involving the raising of a DP from the same base generated structure (Stowell 1978). Moro (1997, 2018) showed that the predicate DPs (including *there* and its equivalents across languages) can be raised along with the subject DPs to the preverbal position from the so-called *Small Clause* (SC) – a structure resulting from merging two DPs (Moro 2000, 2009 Chomsky 2013, Rizzi 2016). In other words, while in Canonical copular sentences the subject DP raises to the preverbal position and the predicative DP stays *in situ* inside the small clause in the postverbal position (4), in the Inverse copular sentences the predicative DP raises to the preverbal position and the subject DP stays *in situ* inside the small clause in the postverbal position (5).

(5) Canonical copular sentence structure

(6) Inverse copular sentence structure

2.1 Asymmetries in copular sentences

These two different representations offer a principled explanation for many asymmetries across languages. Distinguishing between Canonical and Inverse copular sentences is not

always easy or possible (see Jespersen 1924 as cited in Moro 1997). However, agreement and PP/*ne* sub-extraction offer robust diagnostics. For example, verbs invariably agree with the subject DP in Italian (7), regardless of the pre-verbal or post-verbal position, while they invariably agree with the preverbal DP in English (8):

(7) a. le foto   **sono**/*è la causa
       the pictures   are /*is the cause
    b. la causa **sono**/*è le foto
       the cause   are/*is   the pictures
                                        Italian

(8) a. the pictures **are**/*is the cause.
    b. the cause *are/**is** the pictures
                                        English

Extraction is only allowed from the post-verbal DP – the predicate – in Canonical sentences (9), whereas it is not allowed from the post-verbal DP – the subject – in Inverse copular sentences (10).

(9) a. **which riot$_i$** do you think a picture of the wall was **the cause of _$_i$**?
    b. **di quale rivolta$_i$** pensi che una foto del
       of which riot$_i$ do you think that a picture of_the
          muro sia la causa _$_i$?
          wall is the cause _$_i$?

(10) a. *__which wall$_i$__ do you think a cause of the riot was a **picture of _$_i$**?
     b. *__di quale muro$_i$__ pensi che la causa della
        of which wall$_i$ you think that the cause of_the
           rivolta sia **una foto _$_i$**?
           riot   is a picture _$_i$?

3. Experimental evidence supporting the analysis of copular sentences

Before considering the computational side or the proposed structural analysis we investigated whether the human parser is sensitive to the critical distinctions illustrated here. Two experiments are discussed, testing the processing of Canonical vs Inverse copular sentences (first condition) involving the extraction of a wh-element from a DP embedded either under the subject or the predicate (second condition).

Our prediction was that the sensitivity to agreement and to the *argumental* vs. *predicative* role distinction for the two DPs involved would have influenced both the online and the offline performance of native speakers: participants should show an advantage in parsing Canonical copular sentences (vs. Inverse ones), since only

the Canonical configuration allow the extraction from the predicate DP, whereas all the other kinds of extraction – from the subject in Canonical and from both the subject and the predicate in Inverse – should be disallowed (§2.1).

In order to test these hypotheses, we performed (i) a Self-Paced Reading (SPR) experiment with a Sentence Comprehension Task at the end, and (ii) an Acceptability Judgement Task (AJT).

3.1   Material and methods

In both the SPR and AJT the set of stimuli was the same: 128 items (divided in 4 conditions) and 40 fillers, in SPR, and 60 fillers, in AJT per condition (72 items per experiment in SPR, 92 in AJT). The 2x2 design produced four experimental conditions, exemplified in (11):

(11) *Condition 1*:
     Canonical + Extraction from the Subject
*[$_{PP}$ Di quale muro]$_i$ … [$_{DP}$ le **foto** _$_i$]$_a$ sono [$_{SC}$ [_$_a$]
     Of which wall   the pictures are
                 [$_{DP}$ la **causa** [$_{PP}$ della rivolta]]]?
                 the cause   of_the riot?

Condition 2:
Canonical + Extraction from the Predicate
[$_{PP}$Di quale rivolta]$_k$ … [$_{DP}$ le **foto** [$_{PP}$ del muro]]$_a$
     Of which riot   the pictures of_the wall
          sono [$_{SC}$ [_$_a$]   [la **causa** _$_k$]]
          are   the cause?

Condition 3:
Inverse + Extraction from the Subject
*[$_{PP}$ Di quale muro]$_i$…[la **causa** [$_{PP}$ della rivolta]]$_b$
     Of which wall   the cause of_the riot
          sono [$_{SC}$ [le **foto** _$_i$] [_$_b$]]?
          are (=is) the pictures?

Condition 4:
Inverse + Extraction from the Predicate
 *[$_{PP}$ Di quale rivolta]$_k$ … [la **causa** _$_k$]$_b$ sono [$_{SC}$
     Of which riot …   the cause   are (=is)
          [$_{DP}$ le **foto** [$_{PP}$ del muro]] [_$_b$]]?
          the pictures   of_the wall

3.2   Self-Paced Reading

32 native Italian speakers participated in the experiment. Stimuli were composed by questions and by their answers; participants had to read the question word by word and, then, the answer. Finally, they had to judge the appropriateness of the answer.

3.3 Results

Participants showed higher accuracy in answering to comprehension questions when the extraction occurred from the post-verbal DP in Canonical copular sentences – DP *predicate* in Condition 2 – than in Inverse copular sentences – DP *subject* in Condition 3 – while extraction from the Inverse copular constructions induced lower accuracy (-0.41, z=-2.054, p=0.04; Fig. 1). This confirms that the structural asymmetry between referential subjects and predicative DPs has a central role in both the processing and the comprehension of nominal copular sentences. Similarly, Inverse vs Canonical opposition seems relevant since extractions from both sites in the Inverse copular constructions produce lower accurate answers compared to the extraction from the predicate in canonical copulars (coherently with Moro 1997, 2006 that predict the DP in both inverse constructions to be illegal extraction sites).
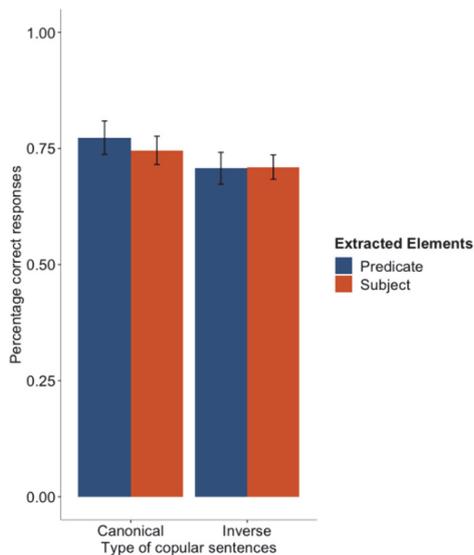
Fig.1 Percentage of correct answers across conditions.

Reading times, on the other hand, revealed a clear difference at the copular region for the two conditions (t=3.37 p=0.002) suggesting a penalty for the Inverse copular constructions compared to the Canonical one. Also at the first DP region the Predicate vs Subject distintion is productively differentialed (t>2 p=0.008) indicating the *la causa* ("the cause") and "*le foto*" ("the pictures") conditions, respectively predicate and subject condition, are perceived as different.

3.4 Acceptability Judgement Task

40 native Italian speakers participated in the experiment. Stimuli were the same than in SPR.

Participants had to rate the acceptability of questions on a scale from 1 to 7.

3.5 Results

The results (fig.2) confirm the previous on-line findings and show that (i) Canonical constructions were more acceptable than Inverse ones and that (ii) among the different types of copular sentences, the ones with an extraction from predicates have higher rates than the ones with extraction from subjects.
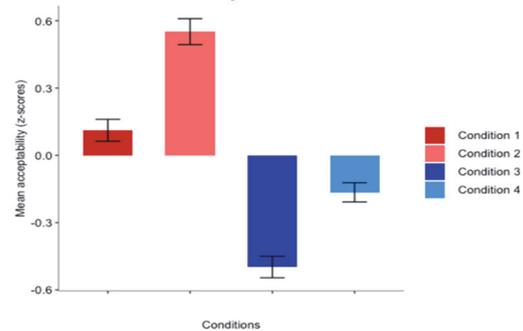
Fig.2 Acceptance rates across conditions.

4. Parsing copular sentences

To evaluate the state-of-the-art of NLP with respect to the contrasts we discussed (Canonical vs Inverse copular sentences) in a configuration where overt agreement disambiguates the critical roles (predicate vs subject), we ran few tests using the following tools:

1. UDpipe (Straka et al 2016)
2. Stanford Parser - English (Chen & Manning 2014)
3. LinguA parser (Attardi, Dell'Orletta 2009)
4. Google Translate (translate.google.com)
5. Google Cloud AI Solutions (cloud.google.com)

We first tested standard Canonical (3.a) and Inverse (3.b) copular constructions, then we tried to assess qualitatively the output analyses provided by these tools with respect to sub-extraction from the predicate in Canonical sentences (9.a-b), here repeated for convenience:

(3)   a. le foto del muro **sono** la causa della rivolta
*the pictures of the wall **are** the cause of the riot*
b. la causa della rivolta **sono** le foto del muro
the cause of the riot   are   the pictures of-the wall
*the cause of the riot **is** the pictures of the wall*

(9) a. **which riot_i** do you think a picture of the wall was **the cause of** __i?
b. **di quale rivolta_i** pensi che una foto del muro sia la causa __i?
*of which riot_i do you think that a picture of the wall is the cause __i?*

4.1 UDpipe

UDPipe Natural Language Processing - Text Annotation interface (Wijffels 2018, Straka et al 2016) provides a handy tool easily integrated in the R environment. Various pre-trained models are available for many languages. We run our analyses using the pre-trained model italian-isdt-ud-2.4-190531. The results of the analysis for both Canonical (10.a) and Inverse (10.b) are simply the same. In fact, not even the basic local dependencies are fully recovered (e.g. det-noun). The analysis of the sub-extraction from predicate in Canonical structures (13.a) is paradoxically less disastrous than the other analyses, but if we try to analyze sub-extraction from the subject of a Canonical construction, we obtain wrong analyses (13.b) (the wh- items is considered an extra argument of *cause*):

(12) a. Canonical copular sentence analysis
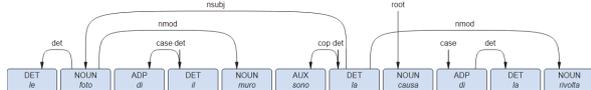
b. Inverse copular sentence analysis
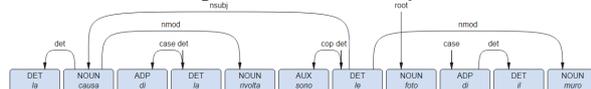
(13) a. sub-extraction from predicate in Canonical configuration
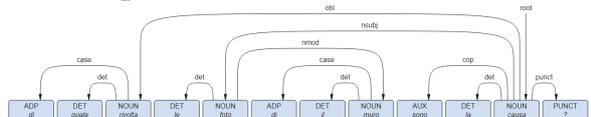
b. sub-extraction from subject in Canonical configuration
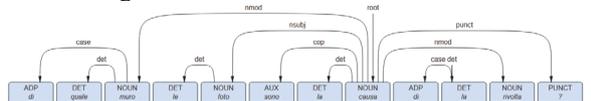
4.2 Stanford Parser

Stanford parser (Chen & Manning 2014) can be considered the state-of-the-art parser for English. Canonical constructions, in fact, gave the opportunity to live up to expectations: the analysis of the canonical copular sentence (14.a) is perfectly in line with the analysis presented in §2-§2.1 (*cause* is identified as predicate and *pictures*

as its subject). Unfortunately, the same analysis is proposed for inverse copular constructions (14.b).

(14) a. Canonical copular sentence analysis
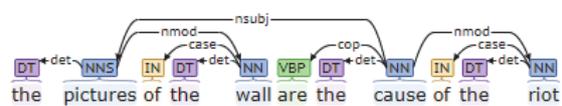
b. Inverse copular sentence analysis
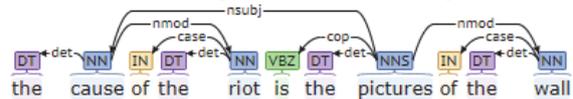
The quality of the analysis for the sub-extraction case confirms every suspicion: the sub-extracted wh-item (*which riot*) is wrongly associated to the matrix predicate (*think*) (15).
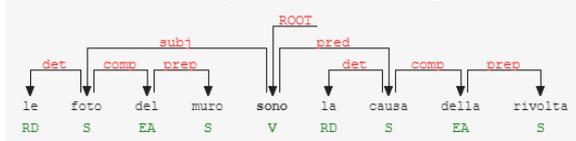
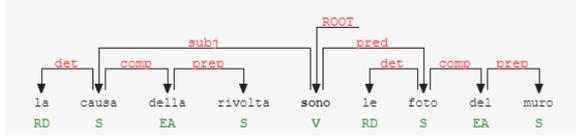(15) sub-extraction from predicate in Canonical configuration
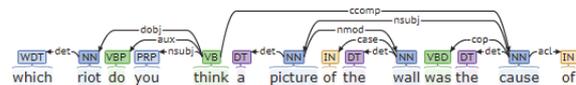
4.3 LinguA

LinguA annotation pipeline (service provided on-line by ItaliaNLP Lab at Istituto di Linguistica Computazionale "Antonio Zampolli" ILC in Pisa) has been used for our tests on Italian, implementing a version of Attardi & Dell'Orletta (2009) parser (currently the state-of-the-art parser for Italian). The analyses of this parser are definitely more precise than the ones proposed by the UDpipe tool, but the symmetric results returned for both Canonical and Inverse copular sentences did not identify either the dependency between the predicate and the subject or their actual role in the structure (16.a-b). The analysis of the extraction, interestingly attempts an interpretation of the *wh-* item as an (extra) argument of the first DP (*le foto [di quale rivolta] (del muro)*). This is a wrong analysis, but it is coherent with the slow-down observed in self-paced reading experiment (§3.3) at the first DP region, though the parser does not make the relevant distinction between subject (17.a) and predicate (17.b) (in this second case, sub-extraction is interpreted as a copula argument).
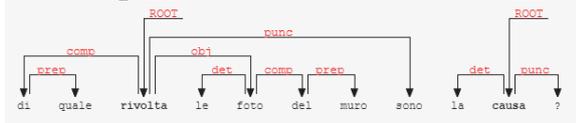
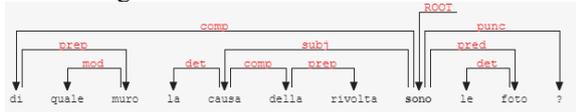(16) a. Canonical copular sentence analysis

b. Inverse copular sentence analysis

(17) a. sub-extraction from predicate in Canonical configuration

b. sub-extraction from subject in Inverse configuration

4.4 Google AI

We finally investigated the Natural Language service – one of the tools provided by Google Cloud AI Solutions API – which returns syntactic representations of sentences (https://cloud.google.com/natural-language/).

While both canonical and inverse copular analyses are equivalent in English to the ones provided by the Stanford Parser (hence partially consistent with our analyses), in Italian, using the Canonical copular sentence '*le intercettazioni$_k$ sono$_k$ la documentazione$_i$*' ('the interceptions are the documentation'), the tool incorrectly analyses the predicate DP *the documentation* as an attribute (fig. 4) (this might be a consistent annotation of all nominal predicates Google adopted, but it is clearly misleading here). Moreover, when it is provided with the Inverse form of the sentence '*la documentazione sono le intercettazioni*' (lett. the documentation are the interceptions; 'The documentation is the interceptions'), the tool incorrectly analyzes the raised predicative DP *the documentation* – singular noun – as the subject, putting it in a wrong agreement relation with the verb (plural form) (Fig. 5). Then, in the end, this parser fails in recognizing the critical difference between Canonical and Inverse copular sentences giving exactly the same analysis for both cases (3.a) and (3.b).
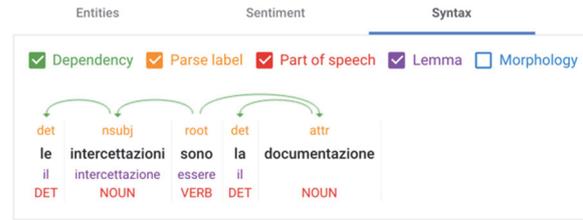

Fig.4 The structural analysis of the Canonical sentence '*le intercettazioni sono la documentazione*' ('The interceptions are the documentation') given by Google Natural Language.
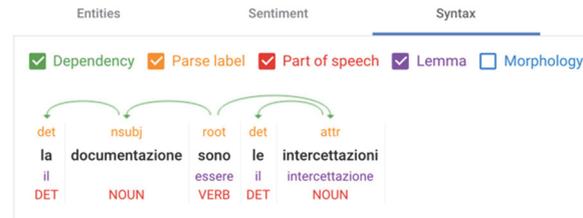

Fig.5 Structural analysis of the Inverse copular sentence '*la documentazione sono le intercettazioni*' (lett. the documentation are the interceptions; 'The documentation is the interceptions') given by Google Natural Language.

4.4 Google Translate

In order to evaluate the impact of these wrong analyses on a practical NLP task, we finally carried out our conclusive experiments on one of the most famous and largely exploited machine translation software: *Google Translate*.

Starting with simple examples, we observed that when the tool is provided with the Italian Inverse copular sentence '*La causa della rivolta sono le foto del muro*' (lett. the cause of the riot are the pictures of the wall; 'The cause of the riot is the pictures of the wall'), it gives the wrong English translation '**The cause of the uprising are the photos of the wall*' (Fig.6), in which the verb does not agree with the pre-verbal DP "*the cause of the uprising*", contrary to what it does in English (as we saw in 7).
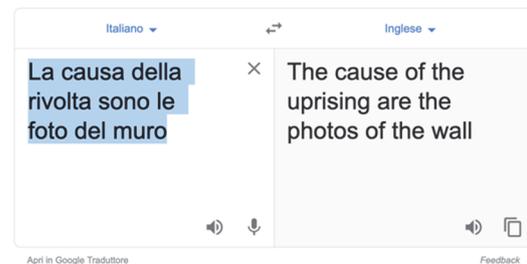

Fig.6 Example from Google translate: https://translate.google.it/?hl=it#view=home&op=translate&sl=auto&tl=en&text=La%20causa%20della%20rivolta%20sono%20le%20foto%20del%20muro

Interestingly, reversing the translation from English to Italian *the cause of the riot is the pictures of the wall* the system correctly produces *la causa della rivolta* **sono** *le immagini del muro* where proper agreement (with the post-verbal subject) is in place. Since the analysis provided by any tool we tested is theoretically inconsistent with this result, we hypothesized that this translation could have been obtained adopting an example-based approach; it was worth then to test if the correct agreement with the post-verbal subject is just an accident (this is a well know prototypical sentence, widely discussed in literature and it might have been included in the Google Translate training set) or if the analysis is generalized of any possible subject/predicate pair.

A sentence like *la documentazione sono le intercettazioni* (lett. the documentation are the interceptions, that means 'The documentation is the interceptions') would suit our purpose nicely. In the English > Italian direction the correct singular copular agreement is produced ("the documentation is the interceptions") but from Italian to English this time the wrong agreement is obtained, totally ignoring the number of the real post-verbal subject (*the documentation is the interceptions* > *la documentazione* **è** *le intercettazioni*). We concluded then that no deep analysis is attempted so as to distinguish between subject and predicate roles and this turns out to be fatal.

5.   Conclusion

In this paper we demonstrated that nominal copular sentences constitute a clear challenge for the computational analysis since the same string of elements [DP V DP] can have in principle two different syntactic representations (hence two different meanings), depending on which kind of copular sentence is realized (Canonical or Inverse). In this paper, we spotted various glitches in the automatic analyses which in the end led either to significant failures (Google Translate) or to rough structural hypotheses that bluntly ignore the relevant contrasts here discussed. Our empirical study, testing both online and offline the *wh-* PP sub-extraction possibilities from both subject and predicate DPs, shows that native speakers are sensitive with respect to the different structural roles; in addition, they perceive as expected the underlying structural representation of Canonical vs. Inverse copular construction. None of the NLP tools we tested succeeded in providing a full set of coherent analyses, with the exception of the Stanford Parser for English that at least succeeded in analyzing correctly the canonical copular sentences. This analysis was however insufficient in the case of inverse constructions and in case of sub-extraction, confirming that non-local dependencies are critical configurations native speakers are able to parse but machine do not, yet.

Reference

Attardi G., Dell'Orletta F. (2009). Reverse Revision and Linear Tree Combination for Dependency Parsing". In: *NAACL-HLT 2009 – North American Chapter of the Association for Computational Linguistics – Human Language Technologies (Boulder, Colorado, June 2009). Proceedings*, Association for Computational Linguistics, 2009. pp. 261 – 264.

Chen D., C. D. Manning. (2014). A Fast and Accurate Dependency Parser using Neural Networks. Proceedings of EMNLP 2014. pp. 740-750

Chomsky, N., (2013). 'Problems of projection.' Lingua 130:33–49

Chowdhury, S. A., & Zamparelli, R. (2018, August). 'RNN simulations of grammaticality judgments on long-distance dependencies.' In Proceedings of the 27th International Conference on Computational Linguistics (pp. 133-144).

Jespersen, 0., (1924) The Philosophy of Grammar, Allen & Unwin, London.

Moro, A., (1991). The raising of predicates: copula, expletives and existence. MIT Working Papers in Linguistics 15: 119-181.

Moro, A., (1997). The Raising of Predicates. Cambridge: Cambridge UP

Moro, A., (2000). Dynamic Antisymmetry. Linguistic Inquiry Monograph, Series, MIT Press

Moro, A., (2006). 'Copular sentences.' In Everaert, M. & H. van Riemsdijk (eds.), MA. Blackwell Companion to Syntax II, Blackwell, Oxford, 1-23.

Moro, A., (2009). 'Rethinking Symmetry: A Note on Labelling and the EPP.' In La grammatica tra storia e teoria: Scritti in onore di Giorgio Graffi, edited by P. Cotticelli Kurras and A. Tomaselli, 129–31. Alessandria: Edizioni dell'Orso; also at http://www.ledonline.it/snippets/allegati/snippets19 007.pdf.

Moro, A., (2018). 'Copular sentences.' In Everaert, M. & H. van Riemsdijk (eds.), MA. Blackwell Companion to Syntax, Revised edition vol. II, Blackwell, Oxford, 1-23.

Rizzi, L., (2016). 'Labeling, maximality, and the head-phrase distinction.' The Linguistic Review 33, 103–127.

Straka, M., Hajic, J., & Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the tenth international conference on language resources and evaluation* (LREC 2016) (pp. 4290-4297).

Stowell, T., (1978). 'What was there before there was there.' In D. Farkas et al., eds., Papers from the Fourteenth Regional Meeting, Chicago Linguistic Society. Chicago Linguistic Society, University of Chicago.

Wijffels, J. (2018). udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the ,UDPipe ',NLP 'Toolkit. R package version 0.5.