Elliptic Constructions: Spotting Patterns in UD Treebanks

Kira Droganova and Daniel Zeman
ÚFAL, Faculty of Mathematics and Physics, Charles University
Malostranské náměstí 25, CZ-11800 Praha, Czechia
{droganova|zeman}@ufal.mff.cuni.cz

Abstract

The goal of this paper is to survey annotation of ellipsis in Universal Dependencies (UD) 2.0 treebanks. In the long term, knowing the types and frequencies of elliptical constructions is important for parsing experiments focused on ellipsis, which was also our original motivation. However, the current state of annotation is still far from perfect, and thus the main outcome of the present study is a description of errors and inconsistencies; we hope that it will help improve the future releases.

1 Introduction

Elliptic constructions (ellipsis) are linguistic phenomena which refer to the omission of a word or several words from a sentence. The meaning of the omitted words, however, can be understood in the context of the remaining elements. For instance, in the sentence “John gave a flower to Mary and [he gave] a book to his son” (an example from (Hajič et al., 2015)) the second predicate and its subject are omitted because of ellipsis. From the syntactic point of view, this significantly alters the sentence structure.

Ellipsis exists in the majority of languages (Merchant, 2001a) and thus deserves careful attention in theoretical and empirical studies, and with regard to NLP applications. The most difficult types of ellipsis (which are the focus of this paper) tend to be rare in comparison to other grammatical patterns, which makes them hard to learn and recognize by parsers. The parsers’ ability to recognize elliptic constructions also heavily depends on the annotation scheme used in a particular corpus: some annotation schemes make ellipsis more visible and identifiable than others.

There is a number of previous dependency analyses of ellipsis. (Mel’čuk, 1988) proposed to use a node labeled as elided, for instance, in the sentence “Alan went to Paris and Leo to Coruña” (an example from (Polguère and others, 2009)), the second verb is marked as elided and thus is invisible. (Lombardo and Lesmo, 1998) used non-lexical nodes and so called non-primitive dependency rules to express gapped coordination. (Osborne et al., 2012) introduced the catena concept and described the elided material of ellipsis mechanisms in terms of catena. (Kahane, 1997) proposed “bubble trees” for gapped coordination.

In this paper we will focus on the basic representation of Universal Dependencies (UD) (Nivre et al., 2016), in which most types of ellipsis are solved by dependent promotion and thus are invisible, i.e., not explicitly annotated as ellipsis; the only exception is missing predicate with multiple overt dependents (orphans). Section 2 gives a brief overview of common ellipsis types and their analysis in UD.

2 Classification

According to (Teselets, 2011), a single rule that motivates elliptical constructions cannot be defined even within one language. The UD guidelines define the following set of rules that proposes a solution for the representation of elliptic constructions:

- If the elided element has no overt dependents, no special relation is required;
- If the elided element has overt dependents, one of those dependents is promoted to take the role of the head;

Following are examples of constructions solved this way:

2.1 Ellipsis in Nominals

When the head noun of a noun phrase is elided (Corver and van Koppen, 2009), according to the
Figure 1: Slovak: An example of adjective coordination, which semantically corresponds to coordination of two full nominals (slovenskej [vedeckej verejnosti] i medzinárodnej vedeckej verejnosti), but the UD approach is to analyze it just as coordinate modifiers. While we consider this approach correct, note that promoting the adjective slovenskej to the head position of the first nominal phrase would lead to a different result: the noun verejnosti would be connected to slovenskej as a conjunct, as shown by the dotted relations below the text.

UD guidelines, one of the orphaned dependents should be promoted to the head position and the other dependents (if any) are attached via the same relations that would be used with the elided head. As a result, there are no means to detect this type of ellipsis in the data (except for unusual POS tag-dependency combinations, such as an adjective serving as a subject).

Coordination of adjectival modifiers can be seen as a special case of an elided noun; however, in this case the usual approach in UD is to just coordinate the adjectives (Figure 1).

2.2 Comparative Deletion
Ellipsis occurs commonly in the complement clause of comparative constructions: in “He plays better drunk than sober,” the full meaning is actually “He plays better [when he is] drunk than [how he plays when he is] sober.”

Here, too, “sober” is promoted all the way up to the head of the adverbial clause that modifies “better”. The relation between the two adjectives is still clausal (advcl); together with the missing subject and copula, these are indirect signs that betray the ellipsis. However, there is no explicit annotation of it.

2.3 Sluicing
Sluicing refers to reduced interrogative clauses, often to a bare interrogative word (Merchant, 2001b). In the following example from UD English, the content in brackets is understandable from the previous sentence: “It’s easy to understand why [the cats refused to eat it].”

Following the UD promotion rules, “why” should be promoted to the head position of the elided complement clause and attached to “understand” via the ccomp relation. (As a matter of fact, it is currently attached as advmod, which we think is an error.)

2.4 VP Ellipsis and Pseudogapping
If a non-finite verb phrase has been elided but a finite auxiliary verb has not, the auxiliary is promoted. Such constructions are called VP-ellipsis (Johnson, 2001) and pseudogapping (Lasnik, 1999). Like with elided nominals, promoting the auxiliary makes these types difficult to identify in the treebank (but see Figure 6 for a counterexample.) Note that the same applies to clauses with non-verbal predicates where the predicate is elided and only copula remains (and is promoted): “John is not smart but Mary is.”

2.5 Gapping and Stripping
Gapping means that the entire predicate is elided, including auxiliary verbs; however, two or more arguments or adjuncts (“orphans”) are overtly expressed\(^1\) (Johnson, 2001; Johnson, 2009; Sag, 1976).

\(^1\)Note that the v2 guidelines mistakenly required the orphans to be core dependents. We argue and demonstrate that the same situation can be caused also by oblique arguments or adjuncts.
In the UD v1 guidelines, the remnant relation was used “to reconstruct predicational or verbal material in the case of gapping or stripping” (de Marneffe et al., 2014); see Figure 2. Practical application showed that such treatment of elliptic constructions has several disadvantages:

- The remnant relation does not produce a clear representation if the second clause contains additional modifiers of the elided predicate;
- The antecedent of the remnant may not exist in the same sentence;
- The annotation style generates many non-projective and parallel structures, thus reducing parsing quality (Nivre and Nilsson, 2005).

The orphan relation is introduced to specify ellipsis more transparently\(^2\) in the UD guidelines v2 (Figure 3). One of the orphaned dependents is promoted and the others are attached to it via the orphan relation. An obliqueness hierarchy is defined, inspired by (Pollard and Sag, 1994);\(^3\) the dependent higher in the hierarchy is promoted. The orphan relation is the only explicit annotation of ellipsis in the basic representation of UD, i.e. only constructions of this type can be easily identified in the data.

UD v2 also defines an enhanced representation where the elided material can be reconstructed using empty nodes (Figure 4). Such representation is currently available only in three treebanks and we do not investigate it further in the present work. Therefore we will focus on the orphan relation in the rest of the paper.

Even more radical reduction is stripping (Hankamer and Sag, 1976) where only one argument remains, assuming that the rest would be identical to

---

\(^2\)http://universaldependencies.org/u/overview/specific-syntax.html#ellipsis

\(^3\)nsubj > obj > iobj > obl > advmod > csubj > xcomp > ccomp > advcl
the previous clause. However, the orphaned argument is usually accompanied at least by an adverb like “too” or “not”. This puts stripping in a gray zone that is not clearly delimited in the UD guidelines. Either we treat the adverb as just a connecting function word, and we attach it to the promoted argument as cc or advmod. Or we treat it as gapping, i.e. the relation is orphan (Figure 9). We cannot quantify the two approaches but both have been observed in the treebanks.

3 Ellipsis in Numbers

Table 1 summarizes the statistics of elliptical constructions in the UD 2.0 treebanks (Nivre et al., 2017). The treebanks are sorted by the last column, which shows the ratio of orphan relations to the total number of nodes in the treebank. 41 treebanks have at least 1 orphan relation in the data, but only 12 treebanks have more than 100 sentences with orphans. Most treebanks have less than 1 orphan per 10,000 nodes, but several treebanks are significantly higher, peaking with the PROIEL treebank of Ancient Greek, which has an orphan in every 500 nodes (Figure 5 shows an example from that treebank).

The number of treebanks which mark elliptic constructions explicitly has doubled since UD release 1.4 (Table 2). However, 29 treebanks from UD 2.0 do not use the orphan relation at all. Some of them are large enough to assume that the studied type of ellipsis actually occurs there but is not annotated properly (we try to address this problem in Section 4.2). Most UD treebanks are conversions of older data annotated under different annotation schemes. If the original scheme does not mark missing predicates somehow, it may not be possible to identify the orphan relations within an automatic conversion procedure.

4 Typical Patterns

Based on the UD guidelines for the orphan relation, one would expect that the most frequent pattern with orphan is coordination of clauses where only the first clause has an overt verbal predicate, while it has been elided from the subsequent conjuncts (clauses). The trees in Figure 3 and Figure 7 are examples of such pattern. However, coordination is not the only possible configuration—Figures 5 and 6 show subordination in a comparative construction. The latter is somewhat less typical in that a copula is promoted but one dependent

| UD Treebank          | Orphans | %    |
|----------------------|---------|------|
| Ancient Greek PROIEL | 701/417 | 0.205% |
| Czech                | 3714/2264 | 0.036% |
| Finnish              | 276/175  | 0.033% |
| Czech CAC            | 1784/1066 | 0.025% |
| Russian SynTagRus    | 2405/838 | 0.02%  |
| Latin ITTB           | 836/607  | 0.014% |
| Romanian             | 66/47    | 0.01%  |
| Greek                | 220/137  | 0.01%  |
| Croatian             | 143/103  | 0.008% |
| Norwegian Bokmaal    | 189/173  | 0.008% |
| Norwegian Nynorsk    | 207/179  | 0.007% |
| Latin PROIEL         | 571/295  | 0.007% |
| Gothic               | 169/96   | 0.005% |
| Old Church Slavonic  | 182/105  | 0.003% |
| Arabic               | 217/72   | 0.003% |
| Slovenian SST        | 28/19    | 0.002% |
| Hungarian            | 64/43    | 0.002% |
| Russian              | 81/66    | 0.002% |
| Catalan              | 12/7     | 0.001% |
| English              | 24/22    | 0.001% |
| Dutch                | 33/12    | 0.001% |
| Swedish              | 44/31    | 0.001% |
| French Sequoia       | 38/29    | 0.001% |
| Slovak               | 110/75   | 0.001% |
| Chinese              | 2/1      | 0.0%   |
| Estonian             | 2/2      | 0.0%   |
| Portuguese           | 7/6      | 0.0%   |
| Italian ParTUT       | 7/7      | 0.0%   |
| Czech CLTT           | 14/11    | 0.0%   |
| Lithuanian           | 3/3      | 0.0%   |
| Coptic               | 2/1      | 0.0%   |
| Belarusian           | 14/7     | 0.0%   |
| Bulgarian            | 3/2      | 0.0%   |
| English ParTUT       | 10/10    | 0.0%   |
| French ParTUT        | 3/3      | 0.0%   |
| Latvian              | 9/8      | 0.0%   |
| Galician TreeGal     | 1/1      | 0.0%   |
| Spanish AnCora       | 29/19    | 0.0%   |
| French               | 3/3      | 0.0%   |
| Swedish LinES        | 4/4      | 0.0%   |
| Italian              | 49/44    | 0.0%   |

Table 1: Statistics on UD v.2.0 treebanks. Orphans: number of orphan nodes/number of sentences. %: the ratio of orphan nodes to all nodes in the treebank.

is still attached as orphan because it complements the elided adjective rather than the whole clause.

The range of dependents that can qualify as or-
Figure 5: Ancient Greek (PROIEL), Herodotus, Histories Book 1: “for they set the first example of war, making an expedition into Asia before the Barbarians made any into Europe.”

Figure 6: English: The copula is is promoted to the position of the elided non-verbal predicate excited.

Figure 7: Portuguese: “O homem fuma entre 11 e 20 cigarros por dia e a mulher, entre 5 e 10.” (“The man smokes between 11 and 20 cigarettes per day, and the woman between 5 and 10.”) The subject of the second clause is promoted and the object is attached to it as orphan. Note that there are other instances of ellipsis, too: entre 5 e 10 [cigarros] (solved by simple promotion of 5), and even the first range, entre 11 e 20 cigarros, in fact stands for entre 11 [cigarros] e 20 cigarros.

phans is rather wide. Core arguments (subjects and objects) are the prototypical cases but oblique arguments or adjuncts (including adverbal modifiers) cannot be excluded (see Figure 8). A special case is the yes-no opposition, rendered in some languages as coordination of a full affirmative verb, and a negative element (without repeating the main verb). Figure 9 demonstrates this on Czech. Note that a similar English sentence would not need the orphan relation: in “they got a meal and I didn’t”, there is an obligatory auxiliary verb in the second part, which gets promoted to the head position.

4.1 Annotation Errors

Ellipsis is a difficult phenomenon, and annotation of ellipsis is a difficult task. Since we are dealing with material missing from the sentence, various annotation styles also miss various bits of information; automatic conversion between annotation styles may have to employ heuristics, and sometimes the correct analysis cannot be obtained without a human in the loop. It is thus not surprising that some of the most common “patterns” we observed in the data are annotation errors. We do not present a complete quantitative evaluation though—we were not able to check all orphans in
In the middle reaches it comes to 30 cm and in the upper reaches it comes to 80 cm. One orphaned adjunct is promoted, the other is attached as orphan.

The negative particle is not considered an auxiliary and is not selected for promotion. Note that if the verb was present, its polarity would be marked by a bound morpheme.

Moreover, the orphan relation should not be used in this situation at all. It is simple coordination of two adjectives, мирните and политически.

The typical error classes are the following:
1. The orphan relation is used instead of conj (Figure 10);
2. Relations are correct, structure is wrong (Figures 11, 12 and 13);
3. The priority of promotion violates the obliqueness hierarchy (Figures 11 and 13);
4. There are two (or more) orphans instead of one, and both are attached to their common ancestor (Figure 14).

all treebanks. However, Table 3 shows some figures for a small number of treebanks. We think that these figures could help contributors to improve their data, but they do not provide a complete overview of the phenomena that are misrepresented in UD treebanks, e.g., the 100% error rate in Spanish AnCora is caused exclusively by erroneous assignment of orphan relation instead of conj relation; the figures for Belarusian and Portuguese cannot be interpreted in a statistically significant way due to small number of sentences containing the orphan relation.
5. The structure is correct but relations are wrong. In particular, some of the treebanks that completely lack orphans fall into this category (Figure 15).

Although we can show examples only from a few treebanks, similar errors can be found in other treebanks, too.

4.2 Search for Missing Orphans

While it is difficult to automatically check whether existing orphan relations are correct, it is even more difficult to identify sentences where an orphan is missing. To prove our hypothesis that the studied type of ellipsis occurs also in treebanks not mentioned in Table 1, we search for the most
5 Conclusion

We have presented the elliptic constructions within the UD 2.0 treebanks. We showed some typical patterns occurring in the data as well as rarely occurring constructions.

The differences in ratio of orphans to treebank size (Table 1) can be explained both by unannotated orphans in some treebanks and by annotation errors ("orphans" instead of "conj") in others. It turned out that the number of annotation errors is rather high which surely reflects the complexity of this linguistic phenomenon.

The current state of the UD annotation w.r.t. ellipsis is insufficient and supports neither parser learning nor (cross-)linguistic studies. While human revisiting of the data is desirable, it is obviously not possible for all the treebanks, and automatic tests/corrections would be highly desirable. We have shown that such tests can at least partially help, and we collected a number of examples, which will hopefully help to improve future heuristics for identifying ellipsis in UD.
Table 2: Statistics on UD v.1.4 treebanks. Remnants: number of remnant nodes/number of sentences. %: the ratio of remnant nodes to all nodes in the treebank.

| UD Treebank            | Remnants | %    |
|------------------------|-----------|------|
| Ancient Greek PROIEL   | 1117/458  | 0.403% |
| Finnish                | 352/175   | 0.047% |
| Romanian               | 128/56    | 0.022% |
| Croatian               | 259/166   | 0.016% |
| Greek                  | 230/149   | 0.011% |
| Latin PROIEL           | 780/344   | 0.011% |
| Gothic                 | 297/120   | 0.009% |
| Norwegian              | 256/230   | 0.008% |
| Hungarian              | 169/68    | 0.007% |
| Old Church Slavonic    | 325/145   | 0.007% |
| English                | 92/54     | 0.004% |
| Russian                | 177/89    | 0.004% |
| Chinese                | 4/1       | 0.0%   |
| Coptic                 | 6/2       | 0.0%   |
| English ESL            | 5/4       | 0.0%   |
| Bulgarian              | 4/3       | 0.0%   |
| Kazakh                 | 22/9      | 0.0%   |
| Galician TreeGal       | 15/10     | 0.0%   |
| French                 | 1/1       | 0.0%   |
| Portuguese Bosque      | 24/11     | 0.0%   |
| Ukrainian              | 6/2       | 0.0%   |

Table 3: Manually assessed error rate in selected treebanks. Err/Sent: number of erroneous sentences/number of sentences with orphans. %: error rate.

| UD Treebank      | Err/Sent | %    |
|------------------|----------|------|
| English          | 1/22     | 4.55% |
| Italian          | 3/44     | 6.82% |
| Belarussian      | 2/7      | 28.6% |
| Portuguese       | 2/6      | 33.3% |
| Russian          | 48/66    | 72.73% |
| Spanish AnCora   | 19/19    | 100.00% |

Acknowledgments

The work was partially supported by the grant 15-10472S of the Czech Science Foundation, and by the GA UK grant 794417.

References

Norbert Corver and Marjo van Koppen. 2009. Let’s focus on noun phrase ellipsis. In Groninger Arbeiten zur germanistischen Linguistik, volume 48, pages 3–26.
Joakim Nivre, Željko Agić, Lars Ahrenberg, and…. 2017. Universal dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague, http://hdl.handle.net/11234/1-1983.

Timothy Osborne, Michael Putnam, and Thomas Groß. 2012. Catenae: Introducing a novel unit of syntactic analysis. Syntax, 15(4):354–396.

Alain Polguère et al. 2009. Dependency in linguistic description, volume 111. John Benjamins Publishing.

Carl Pollard and Ivan A. Sag. 1994. Head-Driven Phrase Structure Grammar. Studies in Contemporary Linguistics. University of Chicago Press.

Ivan Sag. 1976. Deletion and Logical Form. MIT. PhD dissertation.

Yakov Testelets. 2011. Ellipsis in Russian: Theory versus description. In Typology of Morphosyntactic Parameters, pages 1–6, Moscow, Russia. MSUH.