Not all those who wander are lost, not all that appends is *appos*: A preliminary study about Apposition and its impacts on NLP [*]

Isabela Soares-Bastos[1], Luísa Rocha[1], and Cláudia Freitas[1]

Pontifícia Universidade Católica do Rio de Janeiro

Abstract. In this work, we present a preliminary investigation of apposition in Portuguese. We started with a linguistic survey of this syntactic-discursive relation, and followed with a comparison of the computational treatment given to this phenomenon, focusing on the *appos* relation of the Universal Dependencies project. On account of the divergence between the approaches, we made an initial analysis using a parallel corpus of PT-EN, aiming to verify how the UD guidelines apply to the material – that is of the literary genre, a type of text that only recently has gained attention in NLP studies.

Keywords: Computational Linguistics · Natural Language Processing · Parallel Corpora · Apposition · Universal Dependencies ·

1 Introduction

The apposition is an example of a linguistic phenomenon which often little importance is attributed to, but that is of great potential when it comes to information extraction. In Natural Language Processing (NLP), the identification of the explicative apposition contributes to the elaboration of semantic lexicons [10] , information extraction [7], and appositive structures, amongst others, seem to be responsible for detecting characteristics of literary characters [13]. In general terms, the apposition is defined as a term of nominal aspect that joins another in function of explanation. However, that is roughly the only agreement on the matter, or at least in Portuguese. After the (imprecise) definition, grammatical compendia provide us an exhaustive list of types of apposition, of which the classification varies according to each grammar. In this work, we present a preliminary exploration of the appositive relation.

We started conducting a linguistic survey of this syntactic-discursive relation, followed by a comparison with the computational treatment given to the phenomenon, focusing on the *appos* relation of the Universal Dependencies project

[*] Isabela Soares-Bastos was supported by FAPERJ (Undergraduate research scholarship: E-26/260.060/2017) and Luísa Rocha was supported by PUC-Rio (Undergraduate research scholarship).

[1]. Because of the divergence between those approaches, we made an initial analysis in a PT-EN parallel corpus[2] , with the objective to verify (i) how the UD guidelines apply to the material, that is of the literary genre – a type of text that only recently has gained attention in NLP – and (ii) how much of the diversity of structures reported in the Portuguese grammar occur in the original texts and in their translations. Our intention is to, considering the differences in analysis in English and Portuguese, come to an assessment that deals with several cases of apposition in Portuguese, without losing the cross-language goal of the UD project.

2 Apposition in the grammars

2.1 In Portuguese

We looked for definitions and examples of apposition in 7 Portuguese grammars, and each one mentioned it only briefly. All of them agreed about its nominal character, and that the apposition is a noun (or a noun phrase) that adds information or refers to another noun that precedes it. They all also agree that the apposition has the same meaning as the term it refers to, and being able therefore to replace it. From the formal point of view, they agree on the existence of a pause, graphically indicated in writing by a comma:

- A maratonista, modelo de persistência, conseguiu recuperar da operação cirúrgica. [8]
- The marathoner, a model of persistence, recovered from the surgery. [3]

Even though the examples above are quite general, many grammars (, [1], [4], [8]) also mention apposition as the specifying term of generic phrases such as "Rio Amazonas", "Poeta Bilac" and "Cidade de Lisboa" (but not for " Praça de Lisboa", since Lisboa is not a specification of a type of plaza). In these cases, the pause criteria is dismissed, and the feature kept is the semantic correference.

Despite the agreement about the general definitions presented above, the grammars propose different sub-categories for apposition, but they are not always consensual. Our research uncovered 9 different types of apposition: avaliative, classificative, denominative, enumerative, specifying, explicative, identifying, recapitulative and phrasal apposition. Of those, only 3 – enumerative, specifying and explicative – are common to at least 3 authors. However, the limits of these specifications are ill-defined.

[1] http://universaldependencies.org/

[2] A bidirectional parallel corpus "is a type of database with original and translated texts in these two languages that have been linked together sentence by sentence." [5]

[3] Translation provided by the authors.

2.2 In English

According to [2], appositive noun phrases have equivalent status in relation to the noun phrase that they precede, and that is also *head* of the clause. On behalf of this equitable semantic distribution (just as in Portuguese), the head NP and the appositive NP can be inverted without semantic loss.

- the dissident playwright, Vaclav Havel [2]
- Vaclav Havel, the dissident playwright [2]

According to [11], there are 3 types of apposition that can be combined: (i) full or partial – i.e. if either of the sentences in apposition can be omitted without affecting their acceptability, (ii) strict or weak – if they both are from the same syntactic class or not, (iii) nonrestrictive or restrictive - if they refer to the same information unit or not.

The authors also remark the fact that the apposition resembles coordination, as both phenomena handle *"the linking of units of the same rank"*. Therefore, both of the English grammars consulted agree with what is said about apposition in Portuguese when it comes to the semantic criteria.

3 Apposition in the Universal Dependencies framework

Universal Dependencies (UD) is a multilingual project of treebank development, that aims to elaborate an universal inventory of linguistic categories and guidelines to achieve consistent annotation in various corpora, it allows also the use of extensions for specific phenomena in particular languages when necessary.

Apposition (tag *appos*) is placed in the category of Nominal Dependents, alongside with *nmod* and *nummod.* According to the project's official documentation *appos* ("Appositional Modifier") is:

> "An appositional modifier of a noun is a nominal immediately following the first noun that serves to define, modify, name, or describe that noun. It includes parenthesized examples, as well as defining abbreviations in one of these structures. appos is intended to be used between two nominals. In general, modulo punctuation, the two halves of an apposition can be switched. [...]"

3.1 Apposition in UD-Portuguese-Bosque

As indicated in [12], UD-Portuguese-Bosque attributes the tag *appos* for both the explicative and restrictive appositions, benefiting from the original annotation from PALAVRAS [3], despite this decision being outspokenly provisional.

4 Exploring Corpora

As shown so far, there is little consensus regarding the behaviour of apposition: there is no agreement between grammars, or between any of them and the UD guidelines. UD-Portuguese-Bosque, however, recognizes that its positioning is only practical, although it differs from the UD guidelines. We decided, then, to turn to the data – corpora – in order to verify how the definitions already proposed deal with the material. However, instead of UD-Portuguese-Bosque, we opted for using material from a bilingual (PT-EN) parallel corpus, and of the literary genre, for the exploration: the COMPARA corpus [5] [6]. This decision derives from the fact that the literary genre is still very little explored in NLP, whereas it has been gaining some spotlight thanks to the advance of the Digital Humanities. Another relevant matter is being able to contrast how much constructions originally structured as appositive in Portuguese were maintained so in their English counterpart, which could also give us some insight about the particularities of each language when it comes to this phenomenon, even though not absolutely.

We analyzed 100 occurrences of apposition in COMPARA, which was annotated by by the parser PALAVRAS [4]. However, the very notion of right or wrong is still complicated, as there is not yet a common understanding of how apposition should be treated. In many cases, the structures parsed as appositives – and that could, in fact, be in apposition, according to at least one of the consulted grammars - did not fit in any example of *appos* (Table 1). Examining other syntactic relations in the UD guidelines, we stumbled upon *parataxis*, that in various extents resembles certain cases of apposition. In fact, according to [9], the appositive is a type of parenthetical, and parentheticals are mentioned in the parataxis guidelines:

> "The parataxis relation (from Greek for "place side by side") is a relation between a word (often the main predicate of a sentence) and other elements, such as a sentential parenthetical or a clause after a ":" or a ";", placed side by side without any explicit coordination, subordination, or argument relation with the head word. Parataxis is a discourse-like equivalent of coordination, and so usually obeys an iconic ordering. Hence it is normal for the first part of a sentence to be the head and the second part to be the parataxis dependent, regardless of the headedness properties of the language. But things do get more complicated, such as cases of parentheticals, which appear medially."

We considered errors only the cases in which the structures wouldn't be neither apposition or parataxis. Out of the all cases analysed, 42 could be considered either apposition or parataxis. It means that in more than half of the cases (58%) there were errors. The table below show the results.

[4] We considered both the tags *APP* and *N <PRED*

Table 1. Some of the sentences analysed with the data from COMPARA

	Original	Translation
(a)	Era João Romão quem lhes fornecia tudo, tudo, até dinheiro adiantado, quando algum precisava.	João Romão supplied all their needs, even lending them money to tide them over until payday.
(b)	Nada lhes escapava, nem mesmo as escadas dos pedreiros, os cavalos de pau, o banco ou a ferramenta dos marceneiros	They took everything, including bricklayers' ladders, sawhorses, benches, and carpenters' tools.

Error Type	**Amount (%)**	**APP number of cases**	**N>PRED number of cases**
Coordination	26%	10	16
Vocative	5%	3	2
Secondary predicate	9%	0	9
Other erros	18%	5	13

Table 2. Results of the error analysis of a sample of the 100 sentences from COMPARA

5 Concluding Remarks

The investigation we conducted lead us to the following question: do we need the *apposition* category for syntactic analysis, when he have *parataxis*? In Portuguese, some of the sentences used to exemplify apposition could be used to illustrate *parataxis* in UD as well. It is not by chance that the apposition is considered a type of parenthetical which, for their turn, are structures mentioned along the parataxis section.

However, not all that is currently considered *parataxis* in the UD format shares the same characteristics, as the case of reported speech and news articles bylines, such as:

- Washington (CNN)

Those cases, addressed as *parataxis* in the UD format and as *appos* in UD-Portuguese-Bosque, could also be simply nominal modifiers (*nmod*). This is something we intend to investigate further.

Lastly, in the French treebank documentation of the UD guidelines, it is stated that the *appos* relation was replaced by the tags *nmod:appos* (that is used for what in Portuguese would be the specifying apposition), and *conj:appos* (used for the explicative apposition):

- la place Voltaire (*nmod:appos*)
- Sam, mon frère, est arrivé. (*conj:appos*)

Along our study, we had already considered to use *nmod:appos*, a solution that gains strength with the French positioning. And, likewise, perhaps the tag *parataxis:appos* may also be explored in the future. How much of the consistency – of analysis, of annotation and learning – of the syntactic relations is affected by the indiscriminate use of the tag for situations that are, perhaps, distinct? And, the same way, what would happen if what is currently considered to be apposition was merged with parataxis? These are questions we intend to investigate in the future, as part of a study of the impact of the tagsets in learning.

References

[1] Evanildo Bechara. *Moderna gramática portuguesa.* Rio de Janeiro : Nova Fronteira : Lucerna, 2009., 2009.

[2] Douglas Biber et al. *Longman grammar of spoken and written english.* Essex, England : Longman, 1999., 1999.

[3] Eckhard Bick. ≪The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework≫. PhD thesis. Aarhus, Denmark: Aarhus University, 2000.

[4] Lindley Cintra Celso Cunha. *Nova Gramática do Português Contemporâneo.* Lexikon, 2008.

[5] *COMPARA.* Linguateca. URL: `http://www.linguateca.pt/COMPARA`. (accessed: 05.10.2018).

[6] Ana Frankenberg-Garcia and Diana Santos. ≪Introducing COMPARA, the Portuguese-English parallel translation corpus≫. In: *Corpora in Translation Education.* Ed. by Federico Zanettin, Silvia Bernardini, and Dominic Stewart. Manchester: St. Jerome Publishing, 2003, pp. 71–87. URL: `http://www.linguateca.pt/documentos/Frankenberg-GarciaSantos2000.pdf`.

[7] Suemi Higuchi et al. ≪Text Mining for History: first steps on building a large dataset≫. In: *Proceedings of 11th edition of the Language Resources and Evaluation Conference.* Miyazaki, Japan, May 2018.

[8] Ingedore Villaça Koch Mário Vilela. *Gramática Normativa da Língua Portuguesa.* Almedina, 2001.

[9] Mário A. Perini. *Gramática Descritiva do Português.* Editora Ática, 1996.

[10] William Phillips and Ellen Riloff. ≪Exploiting Strong Syntactic Heuristics and Co-training to Learn Semantic Lexicons≫. In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10.* EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 125–132.

[11] Randolph Quirk. *A Comprehensive grammar of the English language.* London : Longman, 1985., 1985.

[12] Alexandre Rademaker et al. ≪Universal Dependencies for Portuguese≫. In: *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling).* Pisa, Italy, Sept. 2017, pp. 197–206.

[13] Diana Santos, Cláudia Freitas, and João Marques Lopes. ≪Ler e estudar a literatura lusófona como parte da literatura mundial: recursos para leitura distante em português≫. In: HD-Rio, Sept. 2018.