DRIPPS: Annotated corpus with discourse relations in Perfect Participial Sentences

António Leal
University of Porto
CLUP
jleal@letras.up.pt

Purificação Silvano
University of Porto
CLUP
msilvano@letras.up.pt

João Cordeiro
University of Beira Interior
INESC TEC
jpaulo@di.ubi.pt

Relevant UniDive working groups: WG1, WG3

Abstract

Discourse Relations (DRels) are meaning relations crucial to analyze discourse structure and to better explain linguistic problems. For that reason, there has been a propagation of small or medium size annotated corpora of different genres (instructive, expository, descriptive, argumentative, narrative; oral, written) and in various languages (individual or parallel): e.g. Penn Discourse Treebank (PDTB) (Prasad et al., 2008), RST Spanish Treebank (RST-ST) (da Cunha et al., 2011), SDRT Annodis French corpus (Afantenos et al., 2012), and Prague Discourse Treebank (Rysová et al., 2016). The increasing interest in annotated corpora with DRels stems from the valuable contribution that those may offer to the development of Natural Language Processing (NLP) applications such as automatic summarization and translation, information retrieval, sentiment analysis, and opinion mining ((Webber et al., 2012) for a review of these applications). For Portuguese, presently, the only existing corpus annotated with DRels is a relatively small corpus of spoken discourse (TED-PT) in European Portuguese (EP) (Zeyrek et al., 2018) and the CRPC Discourse Bank (CRPC-DB) (Mendes and Lejeune, 2022), annotated according to the Penn Discourse Treebank (PDTB) scheme. The closest to other varieties is a cross-document annotation of relations established between sentences aimed at summarization proposed by (Cardoso et al., 2011) for Brazilian Portuguese (BP). There is also (Aleixo and Pardo, 2008), which describes the process of annotation of a corpus of 3534 sentences extracted from news and annotated according to Cross-document Structure Theory, and (Collovini et al., 2007), which annotated a corpus of 50 news using Rhetorical Structure Theory (Mann and Thompson, 1987). Regarding Angolan (AP) and Mozambican (MP) varieties, annotated corpora with DRels are non-existent. Currently, the annotation of DRels in most of the corpora relies on a lexically grounded approach – mostly on information conveyed by discourse connectors (conjunctions or connectives, like ‘although’, ’because’, ’as a result of’) –, which implies leaving some discourse segments without annotation, and few adopt a ‘complete discourse coverage’ (Benamara and Taboada, 2015) taking other information sources into account. However, for a complete annotation, it is essential to consider other Discourse Relational Devices (DRDs) (e.g. semantic and syntactic) that mark DRels. Although these DRDs complex the annotation process, their integration leads to improvement of annotation and a more complete and grounded explanation of discourse organization. Moreover, it is compulsory when the structure under analysis is devoid of discourse connectors, as is the case of most sentences with adverbial perfect participial clauses (APC), that is, subordinated clauses which, in Portuguese, have the auxiliary verb ”ter” in the gerund (’tendo’), or, in English, the auxiliary verb to “have” in the -ing form (“having”), followed by the past participle of the main verb (cf. (1)-(2)).

(1) No passado dia 13 de novembro, o antigo avançado brasileiro já tinha sido submetido a uma intervenção cirúrgica aos rins, tendo recebido alta dois dias depois. (from EP corpus)

On November 13, the former Brazilian striker had already undergone kidney surgery, having been discharged two days later.

(2) Having served his country, he became a great believer in the need for change and to stop unnecessary wars. (from British corpus)

In these cases, the speakers must rely on other sources of information to infer the relevant DRels. Identifying these sources in APC is of utmost importance if one wants to build an algorithm to identify DRels in general automatically. They can give clues to the identification of the relevant sources of information in other constructions where discourse markers are also absent. Furthermore, the study of APC shows crosslinguistic and intralinguistic variation. For instance, (Silvano et al., 2021) use this data to show that, in the computation of temporal relations involving APC without connector in English, the most important parameter is the temporo-aspectual information carried by the perfect participle, whereas in Portuguese the key factors are the relative position of both main and subordinated clauses and their aspectual classes. As for intralinguistic variation, (Silvano et al., 2021) show that AP and MP APC are more alike EP APC and that BP is clearly different from other Portuguese varieties in what concerns the preferred temporal meaning (contrary to what is usually assumed in the literature).

The main purpose of this presentation is to introduce a new language resource, DRIPPS, an annotated corpus of DRels in APC sentences in some varieties of Portuguese, European, Brazilian, Angolan, and Mozambican, and British English. Our interest in the aforementioned Portuguese varieties is motivated by the fact that MP and AP lack, not only annotated corpora, but also stabilized norms, so it is important to uncover the differences and similarities between these Portuguese varieties and the ones that have been more studied and analyzed (EP and BP). Besides, contrary to EP and BP, MP and AP are most likely impacted by other African languages typologically different from Portuguese, such as Bantu languages, so the description of these African Portuguese varieties will contribute to bringing to light their particularities regarding both EP and BP. The inclusion of British English (BE) in the corpus is motivated by two types of reasons. From a theoretical linguistic point of view, it is essential to compare languages, especially from different branches/families. From a computational point of view, because English is a well-studied language and for which many computational tools have already been developed, a corpus that contrasts the same construction in English and Portuguese can help to adapt to the Portuguese language tools that were previously designed for English.

The corpus was entirely constructed with data collected from the Web, applying a crawling method specifically designed for that purpose. Several well-known newspaper websites were targeted for each language and variety and relevant sentences were extracted from online news articles. This corpus now comprises 993 APC annotated with DRels with the following Discourse Relational Devices: connectors, ordering of the clauses, temporal relations, tenses, and aspectual types. For the definition of DRels, Semantic annotation framework (SemAF) — Part 8: Semantic relations in discourse, core annotation schema (DR-core) – ISO 24617-8 (ISO, 2016) was used (see also, (Prasad and Bunt, 2015)). The reasons behind this choice for our annotation scheme are two. The first reason concerns interoperability, which is fundamental (Ide and Pustejovsky, 2010) with the rapid expansion of the Semantic Web and Linguistic Linked Data (Chiarcos et al., 2020). The second set of reasons derives from the first and is related to the requirements of interoperable semantic annotation (Bunt, 2015): it is language-independent, general enough to be able to account for specific instances (although in some cases, more granularity is warranted) and it has a well-defined semantics, which can be machine-interpretable. Additionally, a graphical user interface browser has been under development, not only to access and manipulate the corpus but also to allow the activation of specific DRel constraints, thereby selecting specific cases from the data set that can be analyzed separately. Counts and percentages are obtained, and the computation of other powerful statistics will be added in the subsequent versions.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Lydia-Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prevot, Josette Rebeyrolle, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2727–2734, Istanbul, Turkey. European Language Resources Association (ELRA) and Evaluation and Language resources Distribution Agency (ELDA) and Istituto di Linguistica Computazionale (ILC), European Language Resources Association (ELRA).

Priscila Aleixo and Thiago Alexandre Salgueiro Pardo. 2008. Cstnews: um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento cst (cross-document structure theory.

Farah Benamara and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In *Fourth Joint Conference on Lexical and Computational Semantics (* SEM 2015)*, pages 147–152.

Harry Bunt. 2015. On the principles of semantic annotation. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*.

Paula CF Cardoso, Erick G Maziero, Mara Luca Castro Jorge, Eloize MR Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago AS Pardo. 2011. Cstnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105.

Christian Chiarcos, Bettina Klimek, Christian Fäth, Thierry Declerck, and John Philip McCrae. 2020. On the linguistic linked open data infrastructure. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 8–15.

Sandra Collovini, Thiago I Carbonel, Juliana Thiesen Fuchs, Jorge César Coelho, Lúcia Rino, and Renata Vieira. 2007. Summ-it: Um corpus anotado com informaç oes discursivas visandoa sumarizaç ao automática. *Proceedings of TIL*, 121.

Iria da Cunha, Juan-Manuel Torres-Moreno, Gerardo Sierra, Luis-Adrián Cabrera-Diego, Brenda-Gabriela Castro-Rolón, and Juan-Miguel Rolland Bartilotti. 2011. The rst spanish treebank on-line interface. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 698–703.

Nancy Ide and James Pustejovsky. 2010. What does interoperability mean, anyway? toward an operational definition of interoperability for language technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources. Hong Kong, China*.

ISO. 2016. Language resource management-Semantic annotation framework (SemAF) - Part 8 - Semantic relations in discourse,core annotation schema (DR-core). Standard, Geneva, CH.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Amália Mendes and Pierre Lejeune. 2022. Crpc-db a discourse bank for portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, pages 79–89. Springer.

R Prasad, N Dinesh, A Lee, E Miltsakaki, L Robaldo, A Joshi, and BL Webber. 2008. The penn discourse treebank 2.0. in proceedings of the 6th international conference on language resources and evaluation (lrec).

Rashmi Prasad and Harry Bunt. 2015. Semantic relations in discourse: The current state of iso 24617-8. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*.

Magdaléna Rysová, Pavlína Synková, Jiří Mírovskỳ, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Veronika Pavlíková, Jana Zdeňková, et al. 2016. Prague discourse treebank 2.0. *Data/software*.

Purificação Silvano, António Leal, and João Cordeiro. 2021. On adverbial perfect participial clauses in portuguese varieties and british english. *Romance Languages and Linguistic Theory 2018: Selected papers from'Going Romance'32, Utrecht*, 357:263.

Bonnie Webber, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

Deniz Zeyrek, Amália Mendes, and Murathan Kurfali. 2018. Multilingual extension of pdtb-style annotation: The case of ted multilingual discourse bank. In *Proceedings of the 11th Language Resources and Evaluation Conference-LREC'2018*, pages 1913–1919. European Language Resources Association.