UNIDIVE 1st General Meeting (WG1, WG3)

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
jpcc@ubi.pt

Motivation:

- Discourse Relations (DRels) are meaning relations crucial to analyse discourse structure and better explain linguistic problems. **For that reason,** there has been the propagation of small or medium-sized annotated corpora (e.g. Penn Discourse Treebank (PDTB) (Prasad et al., 2008)). **Moreover,** annotated corpora with DRels can be a valuable contribution to developing Natural Language Processing (NLP) applications, such as information retrieval, sentiment analysis, and opinion mining. **Besides,** corpora annotated with DRels in Portuguese are scarce: **European Portuguese** – a small corpus of spoken discourse, TED-PT (Zeyrek et al., 2018); **Brazilian Portuguese** – e.g. CST-news with a cross-document annotation of relations (Cardoso et al., 2011); **Angolan** and **Mozambican Portuguese** – none.
- **Additionally,** the annotation of Drels in most of the corpora currently relies on a lexically grounded approach – mostly on information conveyed by discourse connectors. **However,** it is essential to consider other Discourse Relational Devices (DRDs) (e.g. semantic and syntactic) that mark DRels, **even more** when the structure under analysis is devoid of discourse connectors, as is the case of most **adverbial perfect participial clauses.**

Main purpose:

- present a new language resource, **DRIPPS**, an annotated corpus of discourse relations in sentences with adverbial perfect participial clauses in some varieties of Portuguese (European (EP), Brazilian (BP), Angolan (AP) and Mozambican (MP)) and British English (BE).

The Data:

- Adverbial perfect participial clauses (APC): the auxiliary verb ”ter” in the gerund (”tendo”, ´having’) + the past participle of the main verb.

(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.
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.

- DRIPPS’ analysis is described in **Silvano et al. (2021)**.

Building the corpus:

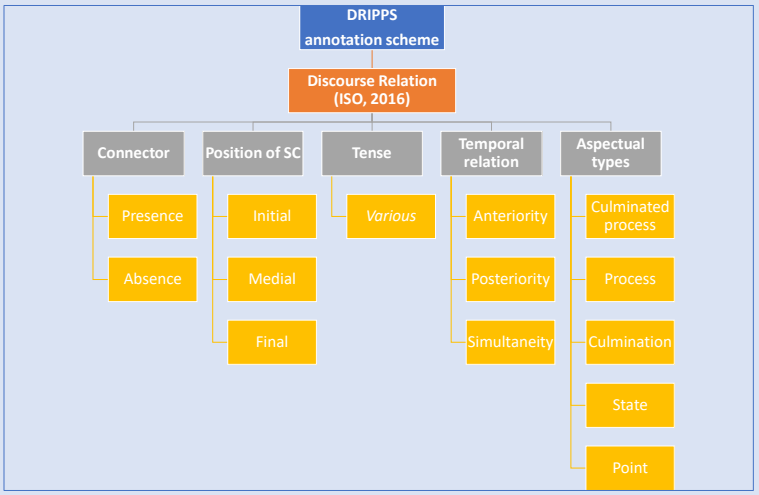
- The corpus was constructed with data collected from well-known newspaper websites by applying a crawling method specifically designed for that purpose.

Variant/language	#Sentcs	#Words	Words/Sentc
European Portuguese	200	7605	38.03
Brazilian Portuguese	193	6734	34.89
Angolan Portuguese	200	7772	38.86
Mozambican Portuguese	200	7262	36.31
British English	200	5715	28.58

Corpus interface browser:

- One sentence per line with its corresponding annotations: Discourse Relation (DR), Semantic Role (SR), Position (POS), Temporal Relation (TR).
- The last column: the sentences, and a given selected sentence is completely visible below in a specific box for that purpose.
- The set of buttons above the table, on the right-hand side: selection of the varieties/languages’ examples to be shown, can be independently activated and deactivated.
- Stats, on the lower side: relevant counts and percentages according to the selections performed in the previous panel of controls. For example, with the path of selections DR → SR → TR, 393 in column (t-3) = total number of records loaded and 108 = the number of cases where DR = cause. The 27.48% in the second line of (t-3) = 108/393.

The annotation framework:



The application:

- https://github.com/johnycordeiro/DRIPPS/

Final remarks:

- **DRIPPS** is a new language resource for Portuguese varieties (a low-recourse language) and for British English that comprises a collection of 993 sentences with adverbial perfect participial clauses with annotations of DRels according to ISO 24617-8 (ISO, 2016), thus ensuring interoperability, and of Discourse Relational Devices intervening in DRels inference (connector, clauses ordering, temporal relation, tenses and aspectual types of both clauses).
- **DRIPPS** comprises an interface browser enabling researchers to better study and explore the DRels phenomena in APC, comparing different Portuguese varieties and even different languages.
- The corpus will continue to be annotated and will be shared with the community so that anyone can effectively analyse and explore DRels.

References
Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., Nunes, M. d. G. V., and Pardo, T. A. (2011). Cstnews- a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In Summarization of News Texts in Brazilian Portuguese. In the Proceedings of the 3rd RST Brazilian Meeting.
ISO (2016). Language resource management-semantic annotation framework (SEMAF) – part 8 – Semantic relations in discourse, core annotation schema (DR-core). Standard, Geneva, CH.
Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The Penn Discourse Treebank 2.0. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May – 1 June 2008, Marrakech, Morocco. ELRA.
Silvano, P., Leal, A., and Cordeiro, J. (2021). On adverbial perfect participial clauses in Portuguese varieties and British English. In S. Baauw, F. Drijkoningen and L. Meron, editors, Current Issues in Linguistic Theory (CILT): Selected Papers from Going Romance 32, Chapter 14. John Benjamins Publishing, Amsterdam.
Zeyrek, D., Mendes, A., and Kurfalı, M. (2018). Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May. European Language Resources Association (ELRA).