LDK 2023

DRIPPS: a Corpus with Discourse Relations in Perfect Participial Sentences

Purificação Silvano
University of Porto and CLUP
Via Panorâmica, s/n
4150-564 Porto, Portugal
msilvano@letras.up.pt

João Cordeiro
University of Beira Interior
Rua Marquês d'Ávila e Bolama,
6201-001 Covilhã, Portugal
INESC TEC, Porto, Portugal
jpcc@ubi.pt

António Leal
University of Porto and CLUP
Via Panorâmica, s/n
4150-564 Porto, Portugal
jleal@letras.up.pt

Sebastião Pais
University of Beira Interior and HULTIG
Rua Marquês d'Ávila e Bolama,
6201-001 Covilhã, Portugal
sebastiao@ubi.pt

Abstract

The main objective of this paper is to introduce a new language resource for some varieties of Portuguese - European, Brazilian, Mozambican, and Angolan - and for British English, called DRIPPS (Discourse Relations In Perfect Participial Sentences). The corpus DRIPPS comprises, at the moment, 993 adverbial perfect participial sentences annotated with Discourse Relations and with the following Discourse Relational Devices: connectors, ordering of the clauses, temporal relations, tenses, and aspectual types. Additionally, an application with a *Graphical User Interface* (GUI) has been developed not only to browse and manipulate the corpus but also to allow the activation of specific Discourse Relation constraints, thereby selecting specific cases from the data set that can be analyzed separately. Besides calculating simple counts and percentages, insightful statistical graphs can be generated and visualized on the fly from the combination of the user-selected constraints and the loaded corpora. The application is pre-loaded with Portuguese and English cases and allows to import/load further cases from different languages/varieties.

1 Introduction

Discourse Relations (DRel) are meaning relations used to describe textual coherence by establishing connections between the different textual segments through meaning functions, crucial to analyze discourse structure and 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 DRel 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 (see Webber et al. (2012) for a review of these applications).

For European Portuguese, the only existing corpora annotated with DRel are the following: a relatively small corpus of spoken discourse (TED-PT) (Zeyrek et al., 2018, 2020; Mendes et al., 2023), and CRPC-DB, a Discourse Bank for Portuguese annotated according to the Penn Discourse Treebank (PDTB) scheme (Mendes and Lejeune, 2022). Regarding other varieties, the closest is CST-news with cross-document annotated relations established between sentences aimed at summarization for Brazilian Portuguese (Cardoso et al., 2011). Aleixo and Pardo (2008) describe the annotation process of this corpus of 3534 sentences extracted from news and annotated according to *Cross-document Structure Theory*. Collovini et al. (2007) annotated a corpus of 50 news texts also in Brazilian Portuguese using *Rhetorical Structure Theory* (Mann and Thompson, 1988). Angolan and Mozambican varieties lack any annotated corpora with DRel.

Currently, the annotation of DRel in many 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 or annotated with implicit relations. Some, nonetheless, adopt a 'complete discourse coverage' (Benamara and Taboada, 2015) taking other information sources into account, like PDTB (Prasad et al., 2008), the American English corpus (Carlson et al., 2001, 2003) annotated with the framework of Rhetorical Structure Theory (Mann and Thompson, 1988) and the Potsdam Commentary Corpus (Stede, 2004), a corpus of German newspaper commentaries also annotated with Rhetorical Structure Theory (Mann and Thompson, 1988), using RST-Tool[1]. For an exhaustive annotation of DRel, it is essential, in addition to discourse connectives, to consider other Discourse Relational Devices[2] (DRD) (e.g. semantic and syntactic) that are pivotal when inferring DRel. The consideration and study of these DRD lead to improved annotation and a more comprehensive and grounded explanation of discourse organization.

Structures without connectives abound in texts, and some have specific syntactic and semantic properties, which may determine the DRel. One such construction is the one with an adverbial perfect participial clause (APC). This type of sentence results from combining two complete propositions, and it can convey inter-propositional values of different types (Móia and Viotti, 2004; Leão, 2018), which can be represented by DRel. Das and Taboada (2018) consider that participial clauses, both with present and past participles, are syntactic signals of certain DRel, that is, they are themselves DRD. However, our study reveals that, although they may signal the existence of a DRel, they allow for a wide array of DRel partly because this construction is mostly devoid of discourse markers. Therefore, the speakers must rely on other sources of information to infer the relevant DRel, such as the tense of the main clause, temporal relations, aspectual type of the situations involved, position of the adverbial perfect participial clause relative to the main clause and the temporal value of the participle. Identifying these sources (or DRD) is essential to better understand how we infer DRel in APC. Moreover, this research can give essential clues to identifying the relevant sources of information in other constructions where discourse markers are also absent. In addition to this, the results of this investigation can also benefit the automatic extraction of DRel.

The primary purpose of this paper is to present a new language resource, DRIPPS, an annotated corpus of discourse relations in sentences with perfect participial clauses in some varieties of Portuguese (European (EP), Brazilian (BP), Angolan (AP) and Mozambican (MP)) and British English (BE), which is the outcome of research that the authors have been developing (Leal, 2011; Silvano et al., 2019, 2021). The option for 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 of utmost importance to uncover the differences and similarities between these Portuguese varieties and the ones that have been studied and analyzed in more depth (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 (e.g. Carvalho and Lucchesi (2016)), so the description of these African Portuguese varieties will contribute to bringing to light their particularities regarding both EP and BP. The inclusion of 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, since English is a well-studied language for which many computational tools have already been developed, a corpus that contrasts the same construction in English and Portuguese can aid in adapting tools designed for English to the specificities of Portuguese.

The following two sections provide a more detailed description of DRIPPS and of an application interface for browsing the corpus. Section 2.1 is dedicated to a brief semantic and syntactic characterization of the data, i.e., sentences with adverbial perfect participial sentences in both languages (Portuguese and British English); Section 2.2 details the process of building the corpus; Section 2.3 lays out the annotation framework; and Section 2.4 presents results of the corpus analysis. Section 3 explains the interface designed to access and work with the corpus. Finally, some concluding remarks and plans for future work are provided in Section 4.

[1] http://www.wagsoft.com/RSTTool/

[2] Term used by TextLink (www.textlink.ii.metu.edu.tr/).

2 DRIPPS corpus

This section describes the Discourse Relations In Perfect Participial Sentences Corpus (DRIPPS), its creation, and the annotation framework. This first version of DRIPPS gathers 993 sentences with adverbial perfect participial clauses in varieties of Portuguese (EP, BP, MP, AP) and British English (BE) annotated with discourse relations (DRel) according to ISO 24617-2:8 (ISO) and relevant discourse relational devices (DRD). More data will gradually be added in the subsequent versions.

2.1 The Data: Adverbial Perfect Participial Sentences

Adverbial perfect participial sentences (APC) (in the Portuguese grammatical tradition, *adverbial gerundive clauses with compound gerund*) are instances of subordinated clauses that, 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) and (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 the EP dataset)
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 BE dataset)

APC have been the object of much research both in Portuguese (mainly for the EP variant, e.g. Leal (2002); Lobo (2003); Móia and Viotti (2004); but also for BP, e.g. Móia and Viotti (2004); (Leão, 2018)), and English (e.g. Quirk et al. (1985); Stump (1985); Kortmann (1995); König (1995)). Overall, APC are described as being introduced, or not, by connectors (subordinating conjunctions or prepositions that function as subordinating conjunctions) and as being able to be placed in an initial and final position regarding their main clause. They are normally featured as conveying temporal interpretations of anteriority or posteriority. Additionally, some studies about the DRel that they may establish indicate that the most frequent are Narration (cf. example (1)), Explanation (cf. example (2)), Result, Background, Elaboration and Concession (Móia and Viotti, 2004; Leal, 2011; Silvano et al., 2019).

Typologically, for European Portuguese, Lobo (2003) divides APC into peripheral clauses, which occur by default in an initial position (with a pause before the main clause) with a temporal meaning of anteriority, and coordinate clauses, which occur only in final position with a temporal meaning of posteriority. However, this proposal is not without problems, as proved by Silvano et al. (2021). The DRIPPS-based analysis carried out by Silvano et al. (2021) reveals that this distinction cannot account for the corpus data since, on the one hand, APC can be positioned initially, finally, and also medially, and, on the other hand, there is not a direct association between the position and the temporal interpretation.

2.2 Corpus Creation through Web Crawling

The corpus of sentences potentially containing APC was entirely constructed with data collected from the *World Wide Web* (Web), applying a crawling method specifically designed for that purpose. A number of well-known newspaper websites were targeted for each language and variety, and relevant sentences were extracted from online news stories. These are well-formed sentences that satisfy specific predefined linguistic patterns provided by the user. We were especially interested in selecting sentences with *adverbial perfect participial* clauses, as described in Section 2.1.

An existing common challenge in the process of selecting well-formed text from web pages is the presence of many “spurious textual segments", like in advertisements, web page structural elements (e.g., menus, sidebars, etc.), and even for news websites. These segments are absolutely unrelated to the news story, with no interest in our study. Another common characteristic of these spurious segments is the lack of an acceptable syntactical structure, even in terms of punctuation marks. Therefore, our text selection method considers these characteristics (more details in Appendix A), selecting only relevant sentences.

The corpus DRIPPS automatically extracted from public online news sources was then manually analyzed, with each sentence classified and annotated by experts from linguistics, as described in Section 2.3. The annotation process adds eight features of information to each selected sentence

related to the DRel, ending up in a data structure as shown in Table 3, as well as in the application interface shown in Figures 3 and 4. Our corpus of 993 adverbial perfect participial sentences, annotated with DRel is stored in conventional and simple CSV format, with one file for each language/variety. These files are directly loaded into the application described in Section 3 and are freely available to the community for research purposes.

Regarding legal issues, it is essential to emphasize that we are not storing whole news texts but only small portions, always keeping the reference to the original source (newspaper URL). The dataset was gathered from publicly available news sources, annotated, and kept only for language research. The decision to resort to online newspapers and not to existing corpora also derives from our intention of studying this structure in comparable, contemporary data.

At the moment, DRIPPS comprises a total of 993 adverbial perfect participial sentences annotated, 793 from four Portuguese varieties and 200 from British English. For Portuguese, DRIPPS has a total 29373 words, representing an average of 37.04 words per sentence. Details on each variety can be observed in Table 1. For the 200 British English sentences, we have a total of 5715 words, giving an average of 28.58 words per sentence.

2.3 Annotation Process

DRel integrate different semantic and pragmatic theories such as *Theory of Discourse Coherence* (Hobbs, 1985), *Rhetorical Structure Theory* (RST) (Mann and Thompson, 1987), or *Segmented Discourse Representation Theory* (SDRT) (Asher and Lascarides, 2003), which differ along several aspects, namely DRel designations, definitions, nature, number, and type of arguments. Bearing in mind, on the one hand, the diversity of these frameworks and, on the other hand, the usefulness of establishing comparisons between annotated corpora from different genres in the same language but also across languages, there have been some efforts to reconcile different proposals of annotation, which have resulted in *Semantic annotation framework (SemAF) – Part 8: Semantic relations in discourse, core annotation schema (DR-core)* - ISO 24617-2:8 (ISO) (see also (Bunt and Prasad, 2016)). ISO 24617-2:8 stipulates an interoperable core-annotation scheme for low-level DRel, i.e., local dependencies. The reasons behind the choice of ISO 24617-2:8 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). It should be noted that, contrary to what Sanders et al. (2021) claim, ISO 24617-2:8 shows that a complete mapping between different sets of DRel proposed within various frameworks is possible. 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[3]) and it has a well-defined semantics, which can be machine-interpretable.

ISO 24617-2:8 provides a set of core DRel of two types, symmetric and asymmetric: while, in the former, the arguments play the same semantic role, in the latter, Arg1 and Arg2 bear relation-specific semantic roles. Figure 1 provides the definitions of the DRel found in our corpus.

Regarding the process of DRel inference, it is widely accepted that the primary sources of information are of two types: linguistic sources (lexicon and compositional semantics) and non-linguistic sources (world knowledge and the cognitive state of the participants) (e.g. Asher and Lascarides (2003)). Although DRel may be implicit, not signalled linguistically, many are explicit, i.e. there is some linguistic marker, be it a word, lexical expression, tense or syntactic structure. These Discourse Relational Devices (DRD) are significant DRel triggers and are studied in many languages (e.g. Das (2014)). In the case of APC, in the absence of a cue phrase to signal the appropriate DRel, the process of inference must depend on other linguistic sources, namely the semantic value of the perfect participle, tense, aspect, mood and modality of the main clause, the presence of negation, or even the mere relative order of both clauses, among other factors. The study of these factors and their relative weight in the overall interpretation of APC has been pursued both for Portuguese and English (for English, e.g. Quirk et al. (1985); Stump (1985); Kortmann (1995), a.o.; for EP, e.g. Leal (2011); Lobo (2003); Silvano et al. (2021); and, for BP,

[3]Despite the fact that "a future part of ISO 24617 is envisaged that will complement this document by providing a complete interoperable annotation scheme for DRel, while also addressing the multilingual dimension of the standard" (ISO), it has not been published so far.

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

Table 1: Corpus statistics.

	DR-core relations	Definition	Semantic Role	
			Arg1	Arg2
Asymmetric	**Cause**	Arg2 is an explanation for Arg1.	result	reason
	Expansion	Arg2 is a situation involving some entity/entities in Arg1, expanding the narrative of which Arg1 is a part, or expanding on the setting relevant for interpreting Arg1. The Arg1 and Arg2 situations are distinct.	narrative	expander
	Asynchrony	Arg1 temporally precedes Arg2.	before	after
	Concession	An expected causal relation between Arg1 and ¬Arg2 is cancelled or denied by Arg2.	expectation raiser	expectation-denier
	Elaboration	Arg1 and Arg2 are the same situation, but Arg2 provides more detail.	broad	specific
	Exemplification	Arg1 is a set of situations; Arg2 is an element of that set.	set	instance
	Manner	Arg2 specifies how Arg1 comes about or occurs.	achievement	means
Symmetric	**Conjunction**	Arg1 and Arg2 bear the same relation to some situation evoked in the discourse, explicitly or implicitly. Their conjunction indicates that they both hold with respect to that situation.		
	Contrast	One or more differences between Arg1 and Arg2 are highlighted with respect to what each predicates as a whole or to some entities they mention.		
	Synchrony	Some degree of temporal overlap exists between Arg1 and Arg2. All forms of overlap are included.		

Figure 1: Definitions of DRel-ISO 24617-2:8 (ISO; Bunt and Prasad, 2016).

Móia and Viotti (2004); Leão (2018). Our annotation scheme includes the most relevant parameters to infer DRel according to the literature. Figure 2 summarizes the framework utilized in annotating DRIPPS.

After designing the annotation scheme, two trained linguists (both EP native speakers with a good command of English) manually annotated a dataset to ensure that the guidelines were well understood. Afterwards, each annotator was assigned a different dataset to be annotated in an Excel spreadsheet. Each line had one example with only one APC. Sentences with two or more APC were duplicated, and each line was dedicated to the analysis of one and only one APC. Regarding the DRel, the annotator had to choose the most prominent DRel whenever there were two possible interpretations. Although sometimes two readings arose, it is a fact that when the writer wrote the sentence, he/she had a specific communicative goal in mind. Whenever the interpretation was not possible due to the lack of a larger context), the example was discharged.

The inter-rater reliability between the annotators was measured with respect to DRel[4], for each variety/language, through *Cohen's Kappa* (Cohen, 1960). Generally, the agreement obtained was significant, as shown in Table 2.

Thus, according to the Landis and Koch (1977) criteria, we can see that we have obtained three *perfect* agreements, one *moderate*, and one *substantial* agreement, shown in the third column from Table 2. The varieties where there was initially some uncertainty among the annotators were Portuguese

[4]The inter-annotator agreement regarding the DRD was not performed because their classification is clear-cut.

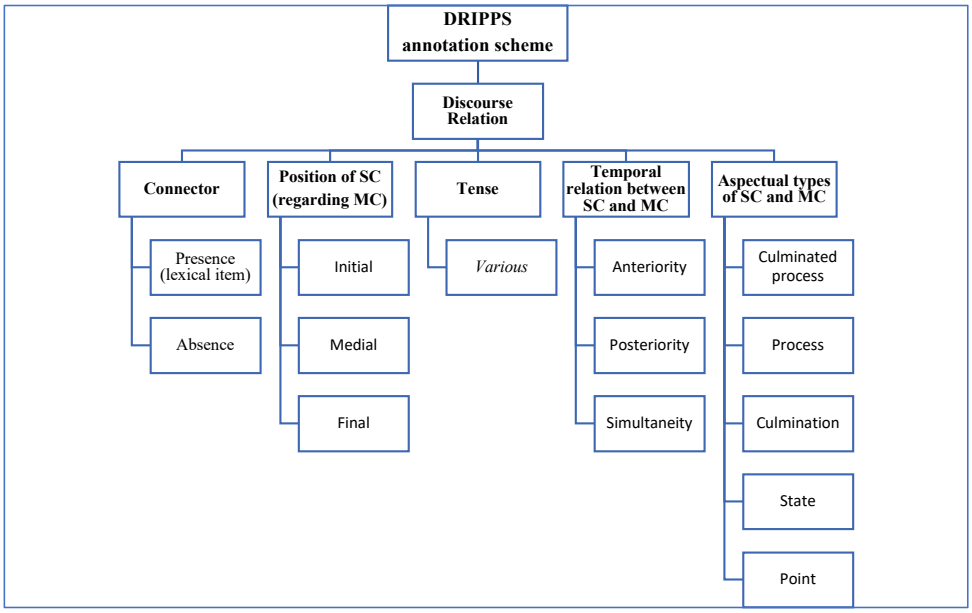

Figure 2: DRIPPS annotation scheme.

Language	*Kappa*	Agreement
PT-Brazil	0.89953	*perfect*
PT-Angola	0.55065	*moderate*
PT-Mozambique	0.67589	*substantial*
PT-Europe	0.95932	*perfect*
EN-Europe	0.88088	*perfect*

Table 2: Annotator agreement measures using *Cohen's Kappa* (Cohen, 1960).

from Angola and Portuguese from Mozambique.

The subsequent step was reaching a consensus regarding the examples of disagreement. The most relevant disagreements between the annotators involved Asynchrony and Conjunction (in AP), Asynchrony and Expansion (in MP) and Expansion (in BP). The annotators discussed the examples and agreed on the accepted DRel.

Table 3 exemplifies the result of manual annotation.

2.4 Some main results of the corpus analysis

From the complete corpus with 4222 sentences in Portuguese and 2635 in English, 993 have already been annotated (EP, AP, MP and BE – 200 sentences each; BP – 193 sentences). This first annotation has already enabled a comprehensive study of the main features annotated in DRIPPS. Silvano et al. (2021) demonstrate that there is crosslinguistic and intralinguistic variation. Since the main objective of this paper is not to present an in-depth contrastive semantic analysis of the data presented in DRIPPS, we refer the reader to Silvano et al. (2021) and present only the main results from the research.

Silvano et al. (2021) conclude from the corpus analysis that, in interpreting temporal relations involving APC without connector in English, the most critical parameter is the temporo-aspectual information given by the perfect participle. In contrast, the key factors in Portuguese are the relative position of both main and subordinated clauses and their aspectual classes. Although there are no absolute restrictions regarding telicity and durativity, aspectual classes of predications are closely intertwined with temporal interpretation as anteriority and posteriority readings tend to be related to telic situations in main and subordinated clauses, whereas simultaneity readings lean on the presence of durative situations in both clauses. In English, by contrast, the combination of aspectual types in both clauses was not a relevant factor, as the anteriority reading is recurrent, irrespective of the aspectual types of both clauses. This is in line with the literature on these structures in English, which points out the anterior orientation of APC.

As for intralinguistic variation, the study also reveals that AP and MP APC are more alike EP APC and that BP is clearly different from other Portuguese varieties in what concerns the main aspects of APC. This finding goes against the idea of an Afro-Brazilian continuum of Portuguese (cf.

Sentence	Pos	TR	Tense MC	ATMC	ATSC	CNT	RR	SR-SC
A PSP do Seixal, no distrito de Setúbal, anunciou nesta terça-feira a detenção de sete pessoas por suspeita de tráfico de droga, tendo sido apreendidas mais de quatro mil doses de droga e 16 mil euros em dinheiro (EP dataset).	Final	Ant	PP	Culm	Culm		asynchrony	before
Até à chegada da troika a Portugal, as despesas com pessoal consumiam sistematicamente 13 % a 14 % do PIB, tendo mesmo atingido o pico de 14,5 % em 2005. (EP dataset)	Final	Simul	PIMP-Ind	St	Culm		expansion	expander
As declarações estão a criar ondas de choque no meio judicial, entre magistrados e advogados, tendo levado o Conselho Superior de Magistratura (CSM) a abrir um inquérito para tirar as insinuações a limpo. (EP dataset)	Final	Post	PresPro	St	Culm		cause	result
Vários profissionais do cinema, inclusive o Exército dos Estados Unidos da América, reagiram a morte de Lee, tendo agradecido o serviço que prestou . (AO dataset)	Final	Simul	PP	Pro	Culm		elaboration	specific
No PSL, que dobrou bancada (de 1 para 2), quem fica fora é Sargento Pereira Júnior, mesmo tendo aumentado sua votação de forma considerável: de 1.267 para 1.530 votos. (BP dataset)	Final	Ant	Pres-Ind	St	CP	mesmo	concession	e-raiser
Segundo o biólogo, a invasão em Moçambique compreende duas vagas: a primeira ocorreu nos fins da década de 60 e início da década de 70, tendo afectado a Ilha da Inhaca. (MZ dataset)	Final	Simul	PP	Pro	Pro		conjunction	
If she was failing, she deserved, after having achieved so much, to be allowed to fail at the polls. (BE dataset)	Medial	Ant	Pst	St	Culm	after	cause	reason

Table 3: Sample of the annotation.

Petter (2009)).

3 The Corpus Interface Application

This section briefly presents the DRIPPS corpus interface application, focusing on the main features implemented so far. The application allows one to load corpora, Portuguese varieties, and British English, in our case, and apply a set of selection constraints to obtain different views and statistics of the data, enabling a whole range of specific corpora analyses and studies. Figure 3 presents the application's main view, where the dataset of annotated sentences from different varieties/languages might be loaded into the main table, the main component of this view. The table presents one sentence per line with its corresponding annotations: *Discourse Relation* (DR), *Semantic Role* (SR), etc. The last column contains the sentences, which are not entirely visible. However, each table's selected sentence is totally visible below in a specific box for that purpose (light yellow colour). The set of buttons above the table, on the right-hand side, allows one to select the varieties/languages' examples to be shown. Each one of these buttons can be independently activated and deactivated, meaning that different sets of varieties/languages can be combined and loaded into the table. In the screenshot from Figure 3, we can see that only the European (EP) and Brazilian Portuguese (BP) varieties are selected. Note that in the table's first column, the prefix of the ID represents the language+variety identification. For instance, the selected example (PTEU197) is from European Portuguese, and the example immediately following is from Brazilian Portuguese. The set of controls (combo boxes) below the table allows one to define DRel constraints to be applied to the table's fields. For example, the configuration presented states that the *discourse relation* (DR) must be *cause*, the *semantic role* (SR) is equal to *reason* and the *temporal relation* (TR) must be of *anteriority* (*Ant*). Different combinations can be set here, and different data examples will be shown accordingly in the table.

The frame of numbers appearing on the lower side of this view, entitled "Stats", shows relevant counts and percentages according to the selections performed in the previous panel of controls. For each new selection, calculations are made, and values are shifted to the right, from `(t)` column toward `(t-3)`. The meaning of these values depends on the path of selections the user decides to follow. For example, here, the path of selections was DR $\rightarrow$ SR $\rightarrow$ TR. Therefore, 393 in column `(t-3)` represents the total number of records loaded (for both varieties), and 108 is the number of cases from these where DR = *cause*. The 27.48% in the second line of `(t-3)` is obtained from $\frac{108}{393}$.

Finally, Figure 4 presents the feature of generating statistical distributions for a given data con-

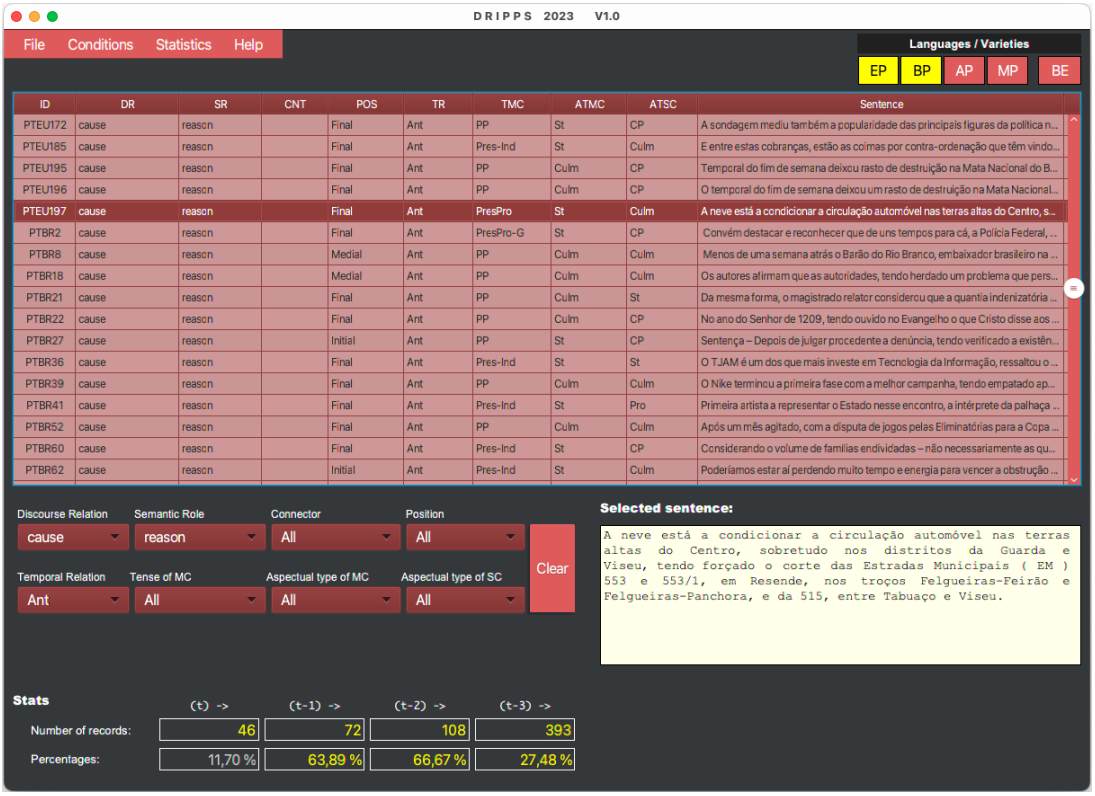

Figure 3: The DRIPPS application to load and explore DRel corpora.

figuration loaded to the applications' table. The data configuration depends on the selected/loaded corpora and the selected constraints applied on the panel. In this particular case, we can see a graph distribution for the *Aspectual type of the SC*, for the *British English* corpus, given that the *Discourse Relation* is set to *cause*. The application allows generating several graphs like this simultaneously and for different data configurations, which enables one, for example, to compare similar phenomena on different corpora.

4 Final Remarks

In this paper, we have introduced a new language resource, DRIPPS, a corpus with an interface browser. This collection of sentences with adverbial perfect participial clauses was extracted from Portuguese varieties (European, Brazilian, Mozambican and Angolan) and British English using a web crawler specially designed and tuned for this task. This first version of DRIPPS gathers 993 APC annotated with DRel according to ISO 24617-2:8 (ISO), thus ensuring interoperability. Moreover, our annotation scheme also includes Discourse Relational Devices intervening in DRel inference, specifically connector, clauses ordering, temporal relation, tense and aspectual types of both clauses. This new language resource comprises an interface browser enabling researchers to better study and explore the DRel phenomena in APC, comparing different Portuguese varieties and even different languages. The corpus will continue to be annotated and shared with the community so anyone can effectively analyze and explore DRel. In fact, the annotated part of DRIPPS has already allowed a wide-range study that highlighted the cross and intralinguistic variation regarding adverbial perfect participial clauses (Silvano et al., 2021). The application that we designed to explore the corpus, due to its versatility, range and the fact that it is user-friendly and intuitive, enables simple but also relevant queries intersecting several parameters.

Although the current state of knowledge about DRel and DRD and their annotation in corpora may be somewhat advanced in several languages, the same cannot be stated for Portuguese, a low-resource language. The research about DRel and the DRD that intervene in the process of inference and are relevant to the creation of automatic annotation methods must be advanced, which is the primary purpose of the current proposal. Manual annotation of these values is the first step to de-

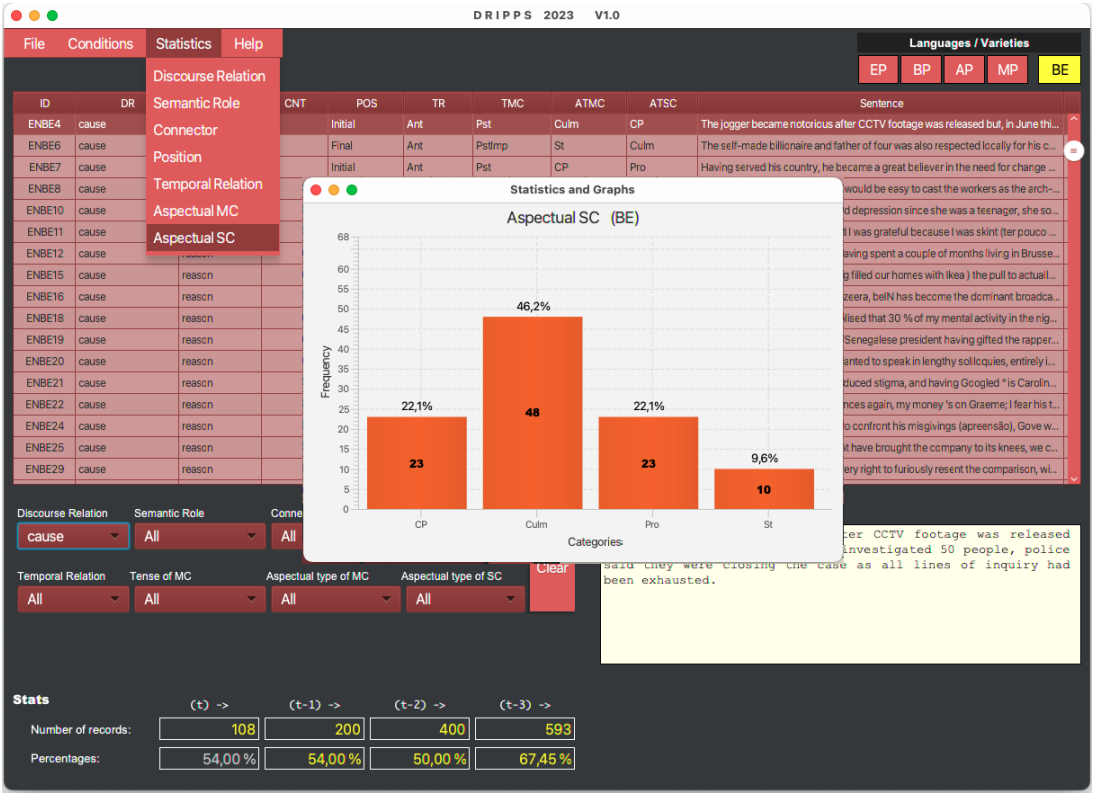

Figure 4: Selecting a statistical graph of the "Aspectual SC" distribution for the BE corpus with Discourse Relation selected on "cause".

velop methods of semi-automatic and automatic extraction of DRel, which we intend to pursue in the future by adapting existing discourse parsers to Portuguese (e.g. Gessler et al. (2021)). Our plans for the future also include extending the annotation to more data of the current varieties/languages. To do so, we will increase the number of annotators, and, "to assess the reliability of an annotation process as a prerequisite for ensuring the correctness of the resulting annotations" (Artstein, 2017), we will not only measure inter-annotator agreement, but also conduct studies about the DRel that cause more disagreement, and the reasons for that disagreement. Lastly, we envisage making the corpus and the interface browser available in the Portulan Clarin infrastructure[5].

Acknowledgements

National funds have funded this research through FCT – Fundação para a Ciência e a Tecnologia, I.P., within the UIDB/00022/2020 project.

[5] https://portulanclarin.net

References

ISO 24617-2: 2016. Language resource management, Part 8: Semantic relations in discourse (DR-core). Standard, Geneva, CH.

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2727–2734, Istanbul, Turkey. European Language Resources Association (ELRA).

Priscila Aleixo and Thiago Pardo. 2008. Cstnews: Um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento cst (cross-document structure theory). Technical Report 326, Universidade de São Paulo.

Ron Artstein. 2017. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, Dordrecht.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, United States.

Farah Benamara and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal.

In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, Colorado. Association for Computational Linguistics.

Harry Bunt. 2015. On the principles of semantic annotation. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. Association for Computational Linguistics.

Harry Bunt and Rashmi Prasad. 2016. ISO DR-core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 45–54.

Paula Cardoso, Erick Maziero, Mara Luca Castro Jorge, Eloize Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago Pardo. 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*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *SIGDIAL Workshop*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory*, pages 85–112. Springer Netherlands, Dordrecht.

Ana Maria Carvalho and Dante Lucchesi. 2016. *Portuguese in contact*, pages 41–55. Wiley Blackwell, Oxford.

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, Marseille, France. European Language Resources Association.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Sandra Collovini, Thiago Carbonel, Juliana Fuchs, Jorge Coelho, Lúcia Rino, and Renata Vieira. 2007. Summ-it: Um corpus anotado com informações discursivas visando à sumarização automática. In *V Workshop em Tecnologia da Informação e da Linguagem Humana – TIL. Proceedings of XXVII Congresso da SBC*, Rio de Janeiro.

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, Hissar, Bulgaria. Association for Computational Linguistics.

Debopam Das. 2014. *Signalling of coherence relations in discourse*. Ph.D. thesis, Arts & Social Sciences: Department of Linguistics.

Debopam Das and Maite Taboada. 2018. RST signalling corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52(1):149.

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jerry R. Hobbs. 1985. On the coherence and structure of discourse. Technical report, CSLI-85-37, Center for the Study of Language and Information.

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*.

Bernard Kortmann. 1995. Adverbial participial clauses in english. In Martin Haspelmath & Ekkehard König, editor, *Converbs in cross-linguistic perspective. Structure and meaning of adverbial verb forms – adverbial participles, gerunds*, page 189–238. Mouton de Gruyter, Berlin.

Ekkehard König. 1995. The meaning of converb constructions. In Martin Haspelmath and Ekkehard König, editors, *Converbs in cross-linguistic perspective. Structure and meaning of adverbial verb forms – adverbial participles, gerunds*, page 1–56. Mouton de Gruyter, Berlin.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

António Leal. 2002. O valor temporal das orações gerundivas em português. In *Actas do XVIII Encontro da Associação Portuguesa de Linguística*, page 455–464, Porto. APL.

António. Leal. 2011. Some semantic aspects of gerundive clauses in european portuguese. *In Cahiers Chronos. From now to eternity. Amsterdam – New York: Editions Rodopi*, 22:85–113.

Rafaella Leão. 2018. *A semântica das construções gerundivas no português europeu e no português do Brasil*. Ph.D. thesis, Universidade do Porto.

Maria Lobo. 2003. *Aspectos da sintaxe das orações subordinadas adverbiais*. Ph.D. thesis, Universidade Nova de Lisboa.

William Mann and Sandra Thompson. 1988. Rethorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, Marina del Rey, CA: Information Sciences Institute.

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*, page 79–89, Berlin, Heidelberg. Springer-Verlag.

Amália Mendes, Deniz Zeyrek, and Giedrė Oleskevicienė. 2023. Explicitness and implicitness of discourse relations in a multilingual discourse bank. *Functions of Language*, 30(1):67–91.

Telmo Móia and Evani Viotti. 2004. Sobre a semântica das orações gerundivas adverbiais. In *Actas do XX Encontro da Associação Portuguesa de Linguística*, page 715–729, Lisboa. Associação Portuguesa de Linguística.

Margarida Petter. 2009. O continuum afro-brasileiro do português. *África-Brasil: caminho da língua portuguesa*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 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*. European Language Resources Association.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Longman, London.

Magdaléna Rysová, Pavlína Synková, Jiří Mírovský, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Veronika Scheller, Jana Zdeňková, and Šárka Zikánová. 2016. Prague discourse treebank 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ted Sanders, Vera Demberg, Jet Hoek, Merel Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2021. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, 17(1):1–71.

Purificação Silvano, António Leal, and João Cordeiro. 2019. Algumas reflexões sobre a classificação de orações gerundivas em português europeu. *Revista da Associação Portuguesa de Linguística*, 5:325–247.

Purificação Silvano, António Leal, and João Cordeiro. 2021. On adverbial perfect participial clauses in portuguese varieties and british english. In Luisa Meroni Sergio Baauw and Frank Drijkoningen, editors, *Current Issues in Linguistic Theory (CILT): Selected Papers from Going Romance 32*, page Chapter 14. John Benjamins Publishing, Amsterdam.

Manfred Stede. 2004. The Potsdam commentary corpus. In *Proceedings of the Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain. Association for Computational Linguistics.

Gregory T. Stump. 1985. *The semantic variability of absolute constructions*. Reidel, Dordrecht.

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

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfali, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. Ted multilingual discourse bank (TED-MDB): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, 54:587–613.

Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 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. European Language Resources Association (ELRA).

A Appendix: Crawling Algorithm

The method we followed to gather sentences from the Web and build our corpus automatically is detailed here in Algorithm 1. One

Algorithm 1 – Web Crawler for Sentence Selection

```
1:  Input: websites, W = {w_1, w_2, ..., w_n}.
2:  sentences ← ∅
3:  for w_i ∈ W do
4:      S_i ← crawlPage(w_i, ∅)
5:      sentences ← sentences ∪ S_i
6:  end for
7:  Store(sentences)
8:
9:  function CRAWLPAGE(url, lnkMem)
10:     text ← selectText(url)
11:     sent ← selectSentences(text)
12:     for u_j ∈ subLinks(url) do
13:         if u_j ∉ lnkMem then
14:             lnkMem ← lnkMem ∪ {u_j}
15:             S_j ← crawlPage(u_j, lnkMem)
16:             sent ← sent ∪ S_j
17:         end if
18:     end for
19:     return sent
20: end function
```

important particularity of this algorithm is the verification of a well-formed sentence (line 10: "selectText($urls$)") during web-page extraction, as well as the satisfaction of the linguistic patterns (line 11: "selectSentences($text$)") pre-defined by the user. As usual, the crawler implements a recursive search method, starting with a given base `URL`, e.g., `www.skynews.com` or `www.expresso.pt`, and then descends into the inner[6] hyperlink hierarchy, avoiding endless loops and repetitive content.

[6] Considering only links pointing to resources within the base `URL`.