DISCOURSE STRUCTURE AND CONTENT ANALYSIS: A COMPUTATIONAL LINGUISTICS' APPROACH

Rui Talhadas
U. Algarve FCHS/INESC ID Lisboa
(rtalhadas@gmail.com)

Nuno Mamede
U. Lisboa IST/INESC ID Lisboa
(nuno.mamede@inescid.pt)

Jorge Baptista
U. Algarve FCHS/INESC ID Lisboa
(jbaptis@ualg.pt)

ABSTRACT

Content analysis is a relevant tool for many human and social sciences, such as Psychology and Sociology, among others. The detection of the structure of the texts is a relevant step in determining how the major content elements are organized. Besides text segmentation into paragraphs, sentences, and clauses, the use of discourse connectors is a fundamental element for the structuring of a text. These connectors include conjunctions and conjunctive adverbs, and they make explicit the meaning relations between sentences forming a text. In this paper, we illustrate a method for capturing the major components of texts and their explicit organization. For evaluation, the method is applied to discourse parsing but it could also be applied to many tasks of content analysis. This interdisciplinary method bridges topics from linguistics and computational linguistics, with possible uses in several areas of social sciences, where content analysis and discourse structure may be relevant.

Keywords: Content Analysis, Text/Discourse Parsing, Discourse Connectors, Portuguese.

JEL Classification: Z00.

1. INTRODUCTION

In Linguistics, Discourse Analysis deals with the higher levels of language encoding, namely with the way texts are structured to adequately perform their communicative goals. One can trace back the modern studies in the field to the seminal work of Zellig S. Harris (1952), in a structuralist perspective, and subsequent theoretical developments, such as Grice's Maxims (Grice, 1975) or the SystemicFunctional Theory (Halliday and Hasan, 1976), some of them more or less influenced by the Philosophy of Language of Wittgenstein (1955).

Content analysis (CA) is an 'umbrella term' that can be described as a set of research procedures and methods, with varying degrees of formalisation, that can be applied to texts in a welldefined and reproducible way and transform them in such a way as to enable the retrieval of meaningful information and produce trustworthy inferences (Tipaldo, 2014). It is "a research technique for the objective, systematic and quantitative description of the manifest content of communication" (Berelson, 1952) Developed since the 1950s, any CA methods must assure the repeatability of the procedures, as scientific reelaboration of texts, and, in the words of one of the founders of CA, aim at answering the questions "Who says what, to whom, why, to what extent and with what effect?" (Lasswell, 1948).

The focus of CA can either be the manifest content of the forms of communication, that is, the very texts in their material and objective form; or the latent meaning, deductively deriving the intentions of the authors of the texts. The former is essentially a quantitative approach that relies mostly in the socalled dictionarybased methods, using statistical analysis to model the distribution of linguistic expressions and arrive at interpretativeprone categories; while approaches to the latent meaning perform qualitative analysis in order to elicit the intentions behind texts and their implications.

Irrespective of the approach adopted, Weber (1990:12) alerts that "To make valid inferences from the text, it is

important that the classification procedure be reliable in the sense of being consistent: Different people should code the same text in the same way". Over the years, much effort has been put into research for operative definitions of inter and intracoder reliability (Krippendorff, 2004, p. 413).

Since the advent of computers and the dissemination of texts through the web, CA has been at the centre of many research domains and it is still today an active field of research in the Social and Computer Science Domains and in the Humanities, in general. Any text or corpus of text can be the target of CA procedures: from medical records, to press, from customer reviews to twits and posts in media networks.

Mass media and communication studies from the late years of the 20th century, which have always had an important role in the assessment of public relations programs and public profile, have now turned to Social media analysis and the impact of new mobile devices in communication processes, in areas that are now known as opinion mining and sentiment analysis (Pan and Lee, 2008; Liu, 2012), and that have a strong economic, social and political impact, influencing stakeholders and deciders alike. Information retrieval and text mining techniques now have limitless access to big data, providing insight on how society interacts and reacts to events and policies, with significant societal impact.

The availability of massive quantities of textual contents in machinereadable form, even those contents are in a non structured form as language, requires the application of natural language processing (NLP) tools to retrieve that information from texts and use it in a wide range of applications (Clark *et al.*, 2010). Several applications are automatic summarization and indexation, topic detection and tracking, among others.

In this sense, the use of NLP techniques can aim at discovering the patterns underlying discourse structure and further process textual content beyond simple wordincontext approach. This is the field of Discourse Parsing.

Discourse parsing is the basis of several methods of automatic content analysis (Neuendorf, 2002). On the subject of discourse parsing, several works in the area of computational linguistics have been developed. Nowadays, most projects on *corpus* annotation of discourse relations are based on the Rhetorical Structure Theory framework (Mann and Thompson, 1988), such as RST Discourse Treebank (Marcu, 2000), which consisted on the annotation of around 30 discourse relations over the Wall Street Journal *corpus*. Other projects, like the Penn Discourse TreeBank (Webber and Joshi, 1998), a version of the Penn TreeBank project (Marcus *et al.*, 1993), use lexical information, having been produced with annotations about discourse connectors, namely conjunctions and conjunctive adverbs. These projects have been created for studies on the English language. Discourse parsers have been developed for several languages, including Brazilian Portuguese (Pardo, 2004; Pardo and Nunes, 2008; Maziero *et al.*, 2015). The later is one of the first attempts at a supervised machinelearning classifier for the identification of relations between text units.

In order to build an automatic discourse parsing system, the first task in hand is to build a discourse segmentation tool, irrespective of the set of discourse relations and the theory of discourse that will then be used. Most of these segmentation tools adopt a rulebased approach (Tofiloski *et al.*, 2009) and it hinges on a comprehensive knowledge about the lexical items connecting discourse units (clauses, sentences, paragraphs), that is, the connective words (and multiword expressions) of the language. This approach leads to higher precision when compared to statistical segmenters. The same approach has also been used for Brazilian Portuguese discourse parser DiZer (Pardo, 2004; Pardo and Nunes, 2008).

In this paper we highlight some of the issues raised in the construction of a discourse segmentation tool for European Portuguese. We present a method and perform a preliminary evaluation over a corpus of scientific texts (abstracts) with strong textual cohesiveness and coherence. This is the first step towards the integration of this prototype into a fullyfledged, rulebased and statistical NLP system for Portuguese.

This paper is structured as follows: First, in §2, we present the main linguistic processes and lexical devices involved in the structuring of discourse by way of the socalled connectors. Then, in §3, we present available linguistic resources and tools for natural language processing of Portuguese texts, in order to present a strategy for capturing the discursive structure of a text. In the next Section, we present the methods, including the *corpus* used on the preliminary experiments. A detailed analysis of several issues found at this initial stage are then presented and discussed in order to build a roadmap towards an efficient and comprehensive discourse parser of Portuguese.

2. LINGUISTIC DEVICES AND PROCESSES IN DISCURSIVE STRUCTURING OF TEXTS

A *text* is a successful piece of communication when it presents internal *coherence* and *cohesion* (Halliday and Hasan, 1976; on Portuguese, see Mendes, 2013). Texts, particularly written texts, are complex linguistic objects, presenting an internal structure, which must be approached in a manner somewhat different from the analysis of simple, isolated, sentences and clauses. Any utterance has structure, but sequences of sentences resource to certain linguistic devices and processes that are not available for simpler sentences. Furthermore, in a written text, formal (editing) devices such as paragraphs, sections, chapters, etc. help produce structure and organize content. We do not consider these types of devices here, though.

In this paper, we are interested in investigating general lexicallybased linguistic devices and processes, operating in written texts, and yielding discursive structure. This are sometimes referred to in the literature as *transition words* (Writing Center, 2014)[66]. We will use a *corpus* of scientific abstracts, the TCC *corpus* (Pardo & Nunes, 2008), consisting in relatively short texts, often with an argumentative structure and other specific rhetorical devices. Our aim, at this time, is mostly to identify the linguistic regularities and the issues that can be raised in the development of a rulebased discourse parser. Our final goal is to develop such parser and to integrate it at a later stage in the STRING natural language processing chain (Mamede *et al.*, 2012). In a way, this is our first step in moving from the already developed Portuguese grammar for the XIP (AitMokhtar *et al.*, 2002), the parsing module of STRING, which aims at intraclausal syntacticsemantic dependency extraction, and advance towards a transsentential, discourse parsing.

In this paper, we focus on the use of two major types of connective devices: conjunctions (§2.1) and conjunctive adverbs (§2.2). Their function in discourse can be seen a kind of "glue", linking together clauses and sentences of a text, rendering it cohesive and coherent. These are, by no means, the sole type of cohesion devices a cogent discourse is made of. Other processes, such as the relative order of the elements in a clause or the sentences' sequence, the coreference relation between separate, even distant, elements of a text (Mitkov, 2002; Marques, 2014), etc., they all contribute in a very relevant way to the cohesiveness and coherence of a text. Nevertheless, and for the strict purpose of this paper, we will ignore them here.

2.1. Sentences and (sub)clauses

From an informationtheoretical viewpoint (Harris, 1991), *clause* and *sentenceboundaries* are the point in the linguistic stream presenting the least constraints on wordsequences. Though this is a comprehensive linguistic notion, practical issues are raised when mechanically parsing sentences in texts, namely in natural language processing of written texts.

Formally, sentence boundaries within texts are relatively easy to determine, being signalled by the use of initial uppercase and specific separators (stop <.>, semicolon <;>, colon <:>, question/exclamation mark <?!>). This depends on the language: some languages do not have such (ortho)graphic devices (Thai), while others have special characters to signal the onset and the end of a sentence (e.g. Spanish ¿? and ¡!). For all practical purposes, we ignore all these sentencesplitting issues in this paper and deem all sentences and paragraphs to be correctly segmented.

Once sentences have been identified as text units, the underlying subunits require a more sophisticated approach. This involves the concept of *clause*, a subsentential unit of sentences, and the corelated concept of *conjunction*. A conjunction is a major partofspeech that can be defined as a category of words joining clauses together within a sentence. *Clauses* can thus be defined as the expression of (at least) one semantic predicate with at least one explicit verb, while *sentences* are sequences of clauses (eventually, only one), related by connective devices, mainly conjunctions. Therefore, sentences formed with a single clause are *simple* sentences, while sentences with two or more clauses are *complex* sentences. Clauses can have different status within sentences: (i) a *main* clause (with a finite tensed form) can be coordinated with another main clause, both having similar or equal status within the sentence (*parataxis*); or (ii) a main clause can have one or more *subordinate* clauses (*hypotaxis*); both processes can be combined in the same sentence, and form complex syntactical structures. Furthermore, there are several types of subordination processes, yielding different types of subclauses (the main types being nominal, adjectival, adverbial, and appositive/parenthetical).

The delimitation of the boundaries of subclauses within sentences and the capture of the semantic relations between them is not a trivial task. In this paper, we adopt an extremely simplified approach: any string introduced by a conjunction (or a conjunctive adverb, see below) is a clause, irrespective of the possibility of having only one or several subclauses (not clearly delimited) within it; any beginning or end of sentence is a clause boundary, as well.

2.2. Conjunctions

Conjunctions convey meaning, and even if a comprehensive and universal semantic classification as not yet been achieved, major types involve the concepts of `<cause>`, `<consequence>`, `<timesequence>`, `<finality/purpose>`, `<comparison>`, etc. For the practical purposes of this paper, we consider that the main traditional semantic categories organizing the set of known conjunctions are sufficient, with some minor adjustments to cover most of the semantic values conjunctions may feature. In fact, even these categories are sometimes difficult to reproduce. Since, in the Harrissian framework, natural language has no external metalanguage (Harris, 1991), the use of the very conjunction may be more informative than any 'artificial' semantic tag, even if this can help to organize semantically similar phenomena.

Conjunctions can be *coordinative* (*mas* 'but') or *subordinative* (*porque* 'because'). In this paper, we lightly address coordination, but we focus rather on subordinative conjunctions introducing (adverbial) subclauses, ignoring other subordination types.

Another important aspect is that conjunctions, both simple and compound (*i.e.* multiword) constitute a finite

[66] http://www.webcitation.org/6FVZvFUW3 (last access: 31/3/2016; all other URL were checked on this date).

set, which can be described extensively. However, to the best of our knowledge, no comprehensive, and universally accepted list of conjunctions is available for Portuguese, especially because of the issues in defining multiword units, as well as the subtle distinction between conjunction and prepositions introducing infinitive clauses. To this paper, we used the (quite extensive) data from STRING system (Mamede *et al.*, 2012), containing about 104 items, along their semantic features. Hence, for example, in the artificial example (1):

(1) *O Pedro fez isso* **enquanto** *a Ana lia o jornal* **mas** *não conseguiu terminar antes dela* **porque** *ela é muito rápida.*
(Pedro did that **while** Ana read the newspaper **but** [he] did not manage to finish before her **because** she is very fast.)

we find a single sentence with several clauses, connected by conjunctions. These clauses can be (manually) delimited (bracketing) and numbered (1 to 4), as shown in (2):

(2) [[*O Pedro fez isso*]$_1$ **enquanto** [*a Ana lia o jornal*]$_2$]$_A$ **mas**
[[*não conseguiu terminar antes dela*]$_3$ **porque** [*ela é muito rápida*]$_4$]$_B$

and the structure between them formalized as in (S), where the specific content of sentences is represented by S_i, as follows:

[S_1 *enquanto* S_2] *mas* [S3 *porque* S4] (S)
mas {[*enquanto* (S_1 , S_2)]$_A$,[*porque* (S3 , S4)]$_B$} (P)

or, alternatively, by the treelike structure (T)[67] shown in Figure 1:

Figure 1: A treelike discourse structure (T) of a sentence with clauses linked by conjunctions

Any of these textual modifications, from the initial discourse (1) to its representations in (2), (S), (P) or (T) is a specific type of *content analysis*, in the sense of Tipaldo (2013:18):

> "Despite the wide variety of options, generally speaking every "content analysis" method implies «a series of transformation procedures, equipped with a different degree of formalization depending on the type of technique used, but which share the scientific reelaboration of the object examined. This means, in short, guaranteeing the repeatability of the method, i.e.: that preset itinerary which, following preestablished procedures (techniques), has led to those results. This path changes consistently depending on the direction imprinted by the interpretative key of the researcher who, at the end of the day, is responsible for the operational decisions made»".

Our aim is to be able to reproduce such analysis mechanically, by way of natural language processing techniques. This could then be used to many languagerelated applications, as in summarization, rhetoric analysis, etc.

In this paper, we adopt the formalism illustrated in (P), as in a secondorder predicate logic. The specifics on the implementation of this formalism are spellout below, in Section §2.3.

2.3. Conjunctive adverbs

Conjunctive adverbs are a hybrid category, halfway between conjunction and adverb. Like other sententialmodifying adverbs, they operate on a sentence. However, their function is to relate that sentence with a previous one. Because

[67] This tree structure was drawn using http://ironcreek.net/phpsyntaxtree/

of this, they are often confused with conjunctions in many grammars. For example, in the following sentence, *porém* (however) is a conjunctive adverb:

> *O Pedro fez isto. A Ana, porém, fez aquilo*
> (Pedro did this. Ana, however, did that)

A set of formal properties distinguishes conjunctive adverbs from other types of adverbs (Molinier and Levrier, 2000). Like other sentencemodifying (as against verbmodifying) adverbs, they have mobility in the sentence and can be fronted to its beginning; they are also outside the scope of the negation of that sentence's main verb:

> *O Pedro fez isto. Porém, a Ana (não) fez aquilo*
> (Pedro did this. However, Ana did (not_do) that)

Besides that, sentencemodifying adverbs can not be extracted by clefting,

A Ana fez aquilo, porém (Ana did that, however)
**Foi porém que a Ana fez aquilo* (It was however that Ana did that)

an operation that can only be used to front sentenceinternal constituents:

A Ana fez aquilo hoje (Ana did that today)
Foi hoje que a Ana fez aquilo (It was today that Ana did that)

Most important, since conjunctive adverbs link the sentence where they occur to the previous sentence, they can not appear in the absolute start of a discourse/utterance, as they require a previous content in order to be accepted and understood.

Exactly like conjunctions, conjunctive adverbs also convey meaning, and the semantic classes they can form are partially the same found for conjunctions proper (`<cause>`, `<consequence>`, etc.) with some further, adverbspecific classes (`<examplifyer>`, `<enumeration>`, etc.).

Because of their particular function, it is not rare to find some of this adverbs used inside a sentence, as if they were conjunctions, complicating issues and giving rise to much ambiguous classifications in traditional grammars:

> *O Pedro fez isto, porém a Ana fez aquilo*
> (Pedro did this, however, Ana did that)

Conversely, otherwise certain clearcut coordinative conjunctions like *mas* 'but' may be used adverbially:

> *O Pedro fez isto mas a Ana fez aquilo, = O Pedro fez isto. Mas a Ana fez aquilo.*
> (Pedro did this but Ana did that)

To the best of our knowledge, besides some partial lists in Costa (2008) and several compound adverbs provided by dictionaries and grammars under the tag of adverbial locutions, the most extensive lists of conjunctive adverbs for Portuguese have been collected and classified by Palma (2009), later revised by Fernandes (2011) in view of disambiguation, and then integrated in the STRING (Mamede *et al.*, 2012) Portuguese grammar and lexicon. This list has undergone constant updating. The current list used for this paper consists of 107 conjunctive adverbs. Most of them were already semantically classified.

Both conjunctions and conjunctive adverbs can be combined in sequences of sentences to produce discourse structure. As mentioned above, these are not the only process language resources to produce cohesion and coherence of discourse, but we define this grammatically shallow devices as the focus of this paper, since they can more easily spotted on the text 'surface'.

2.4. Sentence sequences and the '&' connector

Once all connectors have been parsed and the sentence structure they yield represented in some way, a large number of apparently unrelated sentences remain in most texts. However, if the sequence of sentences is in fact a cohesive and coherent text, they must all be linked by a default connector.

For this situation, Harris (1991) proposes the additive conjunction *and*: on one hand, this is the least constraint

conjunction in any language, whose function is just to put two sentences together with minimal contribution to meaning. Because of the linear sequence in which sentences are ordered in relation to each other in discourse, a temporal (1) and sometimes even causal (2) nexus is often assumed:

(1) *O Pedro leu o jornal, viu um pouco de televisão e telefonou ao filho.*
(Pedro read the newspaper, watched tv for a while and phoned his sun)

(2) *O Pedro foi logo comprar um jornal. Há três dias que não sabia nada de Portugal.*
(Pedro went to buy a newspaper right away. It had been three days since he had got any news from Portugal)

However, several complex factors may vary the semantic relation between consecutive, but otherwise unrelated sentences, foremost the predicates involved in each sentencepair, thus this reconstitution is highly dependent on one's world knowledge.

In this paper, we also assume that any sequence of two sentences (or paragraphs), otherwise unrelated, are nevertheless connected by a dummy coordinative conjunction '&' (= 'and'), but we will abstain from further defining the semantic nexus between those sentences. In the same way, the default connection between paragraphs will be '&&'. Some authors consider this relation a type of ELABORATION (Pardo *et al.*, 2004).

3. LINGUISTIC RESOURCES AND NLP TOOLS FOR PORTUGUESE

In this Section we present the main linguistic resources and natural language processing tools used for the construction of a discourse parser for Portuguese.

3.1. Linguistic resources

The lexicons of conjunctions and conjunctive adverbs of the STRING NLP chain (Mamede *et al.*, 2012) were adapted to the Dela formalism, in order to use them with the Unitex platform (Paumier, 2003, 2016).

In STRING, most of these lexical items are first identified (tokenized and POStagged) in LexMan module (Vicente, 2013) and then syntactically and semantically classified in the XIP parser (AitMokhtar *et al.*, 2002) lexicons. In some cases, the correct tokenization and identification of the POS requires context, so that these tasks are carried out by an intermediate module, RuDriCo (Diniz, 2010; Diniz *et al.*, 2011).

From the initial list, certain entries, particularly prone to parsing errors due to their ambiguity were removed. This is the case of certain simpleword conjunctions (*ao*, caso, *de*, para, *por*, *sem*) that are ambiguous with prepositions, and whose identification requires a more sophisticated parsing tool than Unitex. The same was also done with coordination conjunctions (*e*, *mas*, *nem*, *ou*), since the delimitation of the phrases' and sentences' boundaries connected by coordination is not a trivial task. We also discarded a set of phrases involving pronominal, that is, anaphoric, elements (*além disso*, *por esta razão*, *visto isto*). Not only can these expressions be analysed linguistically, as its correct parsing involves anaphora resolution, which is out of the scope of this paper.

Hence, a final list of 211 entries, 104 conjunctions and 107 conjunctive adverbs, was produced. This small lexicon has been adapted to the Dela format (Courtois, 1990), to be used with the Unitex linguistic development platform. Examples of these conjunctions' lexical entries are shown below:

```
a fim de,.CONJ+subordinate+final

antes que,.CONJ+subordinate+temporal+anterior

depois de,.CONJ+subordinate+temporal+posterior

enquanto,.CONJ+subordinate+temporal+simultaneous

para que,.CONJ+subordinate+final

porque,.CONJ+subordinate+causal

por causa de,.CONJ+subordinate+causal
```

As for conjunctions, a list of conjunctive adverbs was also adapted to be used with the Unitex platform. Here are

some entries of that list:

```
a saber,.ADV+Advconj+appositive
afinal de contas,.ADV+Advconj+consecutive
ainda assim,.ADV+Advconj+concessive
ainda por cima,.ADV+Advconj+additive
antes de mais,.ADV+Advconj+temporal
assim\,,.ADV+Advconj+causal
caso contrário,.ADV+Advconj+conditional
de resto,.ADV+Advconj+concessive
em o entanto,.ADV+Advconj+adversative
isto é,.ADV+Advconj+appositive
ou seja,.ADV+Advconj+appositive
por conseguinte,.ADV+Advconj+consecutive
por enquanto,.ADV+Advconj+temporal
por os vistos,.ADV+Advconj+causal
portanto,.ADV+Advconj+causal
quer dizer,.ADV+Advconj+appositive
```

Using one of the Unitex features, priority was given to these dictionaries, so that these words, when found in a text, are only given the information encoded in our lexicons, while any other information from the system's dictionaries is ignored. This allows us to narrow down the focus of the parser, while accessing the remainder of the information encoded in the system's lexicons. For this paper, since the *corpus* was derived from the Brazilian Portuguese, we also used the lexical resources developed for that variety (Vale and Baptista, 2015 and references therein) and distributed with the Unitex system[68].

3.2. Corpus

For the development of the parser, we used the TCC *corpus*[69] (Pardo and Nunes, 2008). This *corpus* consists of 100 documents with varying length (the shortest with 63 words and the longest with 1,825), 732 paragraphs (average of 7.3 per document), 1,490 sentences (average of 2 per paragraph and 14.9 per document) and 52,644 words (average 71.9 per paragraph, 35.3 words per sentence)[70].

The *corpus* was preprocessed and the texts were splitted with indications of beginning and end of *sentence* (`=s=` and `=cs=`, respectively), beginning and end of *paragraph* (`=p=` and `=cp=`), and beginning and end of *document* (`=doc=` and `=cdoc=`), keeping one document per line (each document is separated by a newline character). Sentence boundaries were defined basically by a full stop followed by uppercase initial[71]. The contractions (*no*=*em*+*o* 'in_the') were also resolved. A manual revision was carried out to ensure correct sentencesplitting and contractionresolving. These transformations on the *corpus* were performed in order to obtain the best possible sentence splitting, while maintaining the possibility of performing a transsentential analysis when processing it with Unitex, otherwise, due to the features of the system, the FST approach would only work within sentence boundaries.

[68] This has been proved to have a significant impact on the number of outofvocabulary (OOV) tokens: Using the European Portuguese resources (Eleutério *et al.* 1995, Ranchhod *et al.* 1999), the number of unknown words was 1,021; while the Brazilian lexicon (Vale and Baptista 2015) only left 635 words without any POS tag.

[69] http://www.icmc.usp.br/pessoas/taspardo/CorpusTCC.zip [20160330]

[70] These countings were made prior to any transformation to the *corpus* and before the 10 sentences randomly selected for the evaluation were removed from the *corpus*. The counts of words (approx. 53,000) and sentences (1,350) presented by Pardo and Nunes (2008) is slightly different, probably due to different tokenization and sentence segmentation criteria.

[71] Colon <:> and semicolon <;> were not treated as sentence boundaries.

The full *corpus*, composed of 100 documents, was divided into two:

- 10% of the documents were randomly removed for evaluation, and;
- the remaining 90 documents were used for the development of the parser.

All calculations mentioned below refer to the development *corpus*. After lexical analysis of the development *corpus* with Unitex, the distribution of the conjunctions and conjunctive adverbs in the *corpus* was obtained. The 10 most frequently occurring items in each class are shown in Table 1.

Table 1: Distribution of most frequently occurring conjunctive adverbs (*AdvConj*) and conjunctions (*Conj*) in the *corpus*

AdvConj	**Count**	**Conj**	**Count**
por exemplo	28	*devido a*	22
ou seja	10	*para que*	21
em_o entanto	9	*além de*	19
por outro lado	8	*quanto*	17
assim,	7	*bem como*	7
portanto	6	*uma vez que*	5
em seguida	3	*nem*	4
isto é	3	*e/ou*	5
por sua vez	2	*apesar de*	5
por um lado	2	*embora*	4

In total, 51 different connectors are used in only 90 texts of the TCC *corpus*, showing the diversity of their use in text. Conjunctions are used the most in these texts (120 instances), though the conjunctive adverbs are very frequent (78 found instances). It this diversity and density, the different combination of them in the same sentence and the different possible positions of the adverbial connectors in the sentence that make their parsing so difficult.

However, the most difficult aspect when identifying connectors is their ambiguity, especially in a tool such as Unitex, with little or no morphosyntactic disambiguation. An example of incorrect POS tagging, resulting from ambiguity, is the output of the following sentence:

> *Segundo Pressman,* **quanto** *mais tarde um erro for encontrado em_o processo de desenvolvimento de software, maior é o custo para correção de esse erro.*
>
> **[quanto, C0Conjsubordcomparative (***Segundo Pressman, #* *mais tarde um erro for encontrado em o processo de desenvolvimento de software, maior é o custo para correção de esse erro***.)]**

In this sentence, *quanto* is part of the proportional (discontinuous) conjunction *quanto mais X, mais Y*. Because the program failed to identify this conjunction correctly, our parser incorrectly classified *quanto* as a comparative conjunction.

Another aspect of ambiguity is the fact that the current resources of Unitex do not produce a POSdisambiguated text, so that when trying to capture clauses, which may be defined as having at least a verb form. Since the text has not been POStagged and disambiguated, one cannot, at this stage, rely on such POS constraint to adequately delimit clauses, as many words are ambiguous between verbs and other POS. Therefore, in this paper, we adopted a very simplistic approach, as far as clause segmentation is concerned, and just considered sentence boundaries, ignoring, for the most part, the sentenceinternal POS tags. This problem will not occur within the STRING fullyfledged NLP system, which is able to produce a fully disambiguated text.

In the next Section, the development of the prototype of the rulebased discourse parser will be explored.

4. A RULEBASED DISCOURSE PARSER PROTOTYPE

The lexicon of conjunctions and conjunctive adverbs of STRING, adapted to the Dela formalism and given a higher priority that the standard Brazilian Portuguese lexicon distributed with the Unitex system, was used to POStag

the development corpus. Then, a simple grammar, consisting of several finitestate transducers (*graphs*) was built. The grammar is divided into four main graphs:

- The first type represents sentences with 2 clauses: a main clause (F_1) and a subordinate clause (F_2), linked by a connector (C): $F_1\ C\ F_2$;
- The second one represents two sentences, linked by a connector, having a sentence boundary (#) between them: $F_1\ \#\ C\ F_2$.
- The third type represents the case when the subordinate clause and the connector is fronted to the beginning of the main clause, and usually this signalled by the use of a comma, separating them; $C\ F_2$**,** F_1 .
- The fourth and last type of construction aims at capturing the recursive nature of coordination and subordination, combining the previous structures after a first parse has been produced.

The graph below exemplifies the FST grammar for the identification of the first type: two clauses connected by a conjunction, within the same sentence.

Figure 2: Excerpt of an FST grammar to identify and classify Conj in sentences with the format:

`=s= F1 Connector F2 =cs=`

The graph above recognizes a main clause (beginning by `=s=`), connected to another clause and within the same sentence (ended by `=cs=`), the two being linked by a causal subordinate conjunction (e.g. *devido a* 'due to'). In the graphs, the grey nodes `F1` and `F2` are subgraphs that represent clauses, the main and the subordinate clause, respectively. These are associated to output variables `$F1$` and `$F2$`, respectively. The conjunction is also associated to an output variable `$Conj$`. A similar graph was built for each type of conjunction. Another set of graphs was also produced for the conjunctive adverbs and, for each type of structure, the complete graph brings them all together. These transducers are applied to the text in *replace* mode. In the output (indicated below the nodes' path), opening/closing delimiters (square brackets: `[` and `]`) are inserted, the connector is moved to the front, leaving the symbol '`#`' in its place, and the clauses are then presented inside brackets in their basic order, that is, first the main and then the subordinate clause. The structure type is indicated in the output by a prefix with the form `C` (for conjunctions) or `S` (for adverbs), and the indexes `0` to `3` (one for each type of construction presented above. When the sequence described by the graph is recognized in the text, the following output is produced:

> *É também difícil saber qual de_os sistemas prontos seria a melhor base para o novo sistema,* **devido a**_*a falta de rigor em a documentação.*
>
> **[devido a, C0Conjsubordcausal (***é também difícil saber qual de_os sistemas prontos seria a melhor base para o novo sistema,* # *a falta de rigor em a documentação.***)]**

The next graph recognizes type #2: a sentence (preceded by `=s=`), a sentence boundary (`=cs= =s=`) and another sentence that begins with a comparative conjunctive adverb. It represents a main clause (F1) subordinating F2, by the use of a comparative conjunctive adverb. The complete graph contains all the other types of conjunctive adverbs. Another graph containing all the types of conjunctions, maintaining the same structure, was produced, with no intention to capture any connector, but in order to ensure none occurred in that position.

Figure 3: Excerpt of an FST grammar to identify and classify AdvConj in sentences with the format:

`=s= F1 =cs= =s= Connector F2 =cs=`

When this sequence is recognized in the corpus, it is given the tag *S1Advconjcomparative*, in the following format:

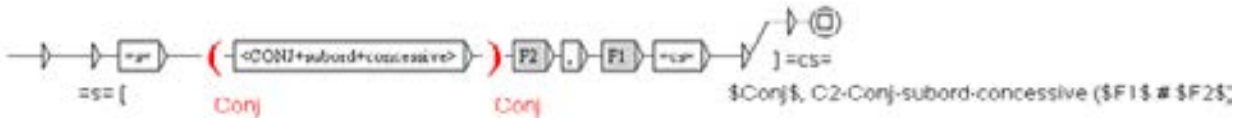
Muitos de_esses sistemas foram feitos baseados em outros sistemas anteriormente por mim desenvolvidos. **Por exemplo** *com base em_um sistema para Controle de Estoque e Emissão de Notas Fiscais de Produtos Agrotóxicos, foi feito um sistema para uma Revendedora de Motocicletas e Peças.*

[Por exemplo, S1Advconjcomparative (*Muitos de_esses sistemas foram feitos baseados em outros sistemas anteriormente por mim desenvolvidos.* **#** *com base em_um sistema para Controle de Estoque e Emissão de Notas Fiscais de Produtos Agrotóxicos, foi feito um sistema para uma Revendedora de Motocicletas e Peças.***)]**

The clauses of the type #2 are identified and classified by the graph below:

Figure 4: Excerpt of an FST grammar to identify and classify *Conj* in sentences with the format:

`=s= Connector F2 F1 =cs=`



It recognizes the case where the subordinate clause is preceded by a conjunction, in this case a concessive, and put before the main clause.

The complete graph contains all the other types of conjunctions. Another graph containing all the types of conjunctive adverbs, maintaining the same structure, was produced, with no intention to capture any connector, but in order to ensure none occurred in that position. When this sequence is recognized in the corpus, it is given the tag *C2Conjsubordconcessive*, in the following format:

Apesar de *ser uma técnica conhecida há algum tempo, os cenários têm ganhado em_os últimos anos grande destaque entre os principais autores em_a área de desenvolvimento de sistemas.*

[Apesar de, C2Conjsubordconcessive (*os cenários têm ganhado em_os últimos anos grande destaque entre os principais autores em_a área de desenvolvimento de sistemas.* **#** *ser uma técnica conhecida há algum tempo.***)]**

One last graph was considered in order to capture the cases where there are three sentences joined by a connector between each pair of sentences/clauses. This graph in applied after all other graphs, because Unitex does not work in a recursive way thus, in order to identify, these sequences, the graph had to take the format acquired by the sentences after being identified with a connector, as exemplified in Figure 5:

Figure 5: Excerpt of an FST grammar to identify and classify AdvConj in sentences with the format:

`F1 (prev. marked sentence) =cs= =s= Connector F2 =cs=`

The complete graph contains all the other types of AdvConj. As in the previous cases, another graph containing all the types of conjunctive adverbs, maintaining the same structure, was produced, with no intention to capture any connector, but in order to ensure none occurred in that position.

When this sequence is recognized in the corpus, it is given the tag *S3Advconjconcessive*, in the following format:

Por causa de_*a complexidade de informações envolvidas, é necessário que a simulação seja apoiada por ferramentas para ser adequadamente realizada.* **No entanto***, uma questão importante que geralmente não é considerada pela simulação é o aspecto de cobertura de_a atividade de teste, através da qual é possível a quantificação da qualidade dessa atividade.*

`[Por causa de, C2Conjsubordcausal` (é necessário que a simulação seja apoiada por ferramentas para ser adequadamente realizada. **`# [em_o entanto, S3Advconjconcessive`** (*a complexidade de informações envolvidas* **#** *uma*

questão importante que geralmente não é considerada por_a simulação é o aspecto de cobertura de_a atividade de teste, através de_a qual é possível a quantificação de_a qualidade dessa atividade.)]

The former graphs are applied to the corpus in different phases:

- first, the FSTs that capture types *#0* and *#2* are applied simultaneously;
- graph for type *#1* is applied after the former because it may overlap some of the cases that are to be tagged by the graph for type *#2*, and;
- graph for type *#3* is the last one to be applied to the corpus, so that the rest of the corpus is already tagged, making it more secure to tag the sequence of three subordinated/coordinated sentences.

The grey boxes in the FSTs represent auxiliary graphs that represent *clauses*, that is, any string of words eventually including some separators, from a sentence boundary to another. The main formal variations include:

- Multiple (i.e.) recursive subordinate clauses modifying a main clause or another subclause
- coordinate clauses
- fronting of a subordinate clause modifying a main clause
- conjunctive adverbs linking two sentences (across a sentence or paragraph boundary)

Table 2: Construction types and their frequency in the training corpus

Construction type	*Conj*	*Advconj*	**Count**
#0 =s= F1 Connector F2 =cs=	137	55	192
#1 =s= F1 =cs= =s= Connector F2 =cs=	0	25	25
#2 =s= Connector F2 F1 =cs=	35	0	35
#3 F1 (prev. marked sentence) =cs= =s= Connector F2 =cs=	0	3	3
Total	172	83	255

Table 3: Distribution of most frequent conjunctive adverbs (*AdvConj*) and conjunctions (*Conj*) in the corpus

AdvConj	**Count**	**Conj**	**Count**
S0Advconjcomparative *por exemplo* (29)	29	C0Conjsubordcomparative *quanto (16), bem como (7), etc.*	29
S0Advconjappositive *ou seja (10), isto é (2), etc.*	13	C0Conjsubordcausal *devido a (15), uma vez que (4), etc.*	24
S1Advconjcomparative *por outro lado (5), por exemplo (3), etx.*	9	C0Conjsubordfinal *uma vez que (12), etc.*	23
S1Advconjcausal *assim (5), portanto (3)*	8	C0Conjsubordadditive *além de (12)*	12
S0Advconjcausal *portanto* (3), *assim (2)*	5	C2Conjsubordcausal *devido a (7), por causa de (2), etc.*	11
S1Advconjconcessive *no entanto (4)*	4	C0Conjsubordtemporal *antes de (2), depois que (2), etc.*	9
S0Advconjtemporal *em primeiro lugar (1), em seguida (1)*	2	C2Conjsubordadditive *Além de (7)*	7
S1Advconjtemporal *em seguida (2)*	2	C2Conjsubordconcessive *embora (3), apesar de (3)*	6
S3Advconjcomparative p*or exemplo (1), por outro lado (1)*	2	C0Conjsubordconcecutive *de forma que (3), de maneira que (2)*	5
S0Advconjconsecutive *nesse sentido (1)*	1	C2Conjsubordfinal *para que (4)*	4

The raw frequencies of the constructions treated in this prototype are shown in Table 2. The data on Table 2 shows that conjunctions are much more used in type #0 (`=s= F1 Connector F2 =cs=`) than in any other type of sentences, being used only in type #0 and #2, which means it is only used to connect clauses that are inside the same sentence, be it in the canonical form or with the main clause after the subordinate.

Conjunctive adverbs, on the other hand, are used in both ways: between sentences (types #1 and #3) and within the same sentence (#0 and #2).

Naturally, the results above represent only the number of matches produced by the system from the development corpus. The error analysis of the evaluation corpus will be addressed in Section 5.

The 10 most frequently occurring tags in each class are shown in Table 3. In both, conjunctive adverbs and conjunctions, the most frequent class of connectors used is the comparative, which represents connectors such as the following:

```
tal como    Conjsubordcomparative

por exemplo Advconjcomparative

bem como    Conjsubordcomparative

assim como  Conjsubordcomparative
```

In the next Section, we present a preliminary evaluation of the prototype discourse parser.

5. EVALUATION

For this paper, and because, to the best of our knowledge, there is no publicly available *corpus* that uses the same segmentation criteria we here adopt, we produced our own evaluation (or reference) *corpus* to assess the performance of our discourse parser prototype. This *corpus* is composed of 10 sentences that were randomly removed from the full TCC *corpus* and manually, and independently annotated, by a linguist. In total, the evaluation *corpus* consists of 10 documents with 103 paragraphs, 215 sentences, and 7.216 words, in total.

The annotation of this sub*corpus* consisted in indicating the relations involving the conjunctions and conjunctive adverbs here considered, as it was described in Subsection 2.2 and henceforward. This annotated sub*corpus* constitutes a *reference* or *golden standard*, against which the system's output is to be compared and evaluated. The annotator did not participate in the development stage of the grammar nor did he have any access to the development sub*corpus*. Conversely, the developer of the grammar ignored the evaluation *corpus*.

The evaluation was performed semiautomatically, and it compared the results from the parser with the manually annotated reference. In the evaluation *corpus*, there is a set of 74 connectors, and the parsing results obtained are the following (Table 4).

Table 4: Results

Results	Count
TP = correctly matched	20
Partially matched	4
FP = incorrectly matched	5
FN = Missed	50

Correctly matched instances correspond to the cases where the tag, and both clauses' boundaries in the output of the system are exactly like what has been encoded in the reference *corpus*. In partial matches, the connector is correct, but one of the clauses is incomplete. Incorrectly matched cases are the ones where there are no connectors present in the sentence but, because of an incorrect processing of the *corpus*, the parser assigns a wrong POStag to some word or expression. Missed instances correspond to the cases where a tag and its parse are present in the reference *corpus* but the parser failed to capture them. The evaluation took into account the standard evaluation metrics of Precision (P), Recall (R), Accuracy (A) and Fmeasure (Resnik and Lin, 2013). These metrics can be defined as follows (TP: true positives; FP: false positives; TN: true negatives; FN: false negatives):

- *Precision* is the sum of all correctly marked cases (TP) divided by the sum of all marked cases (TP+FP):

$$P = \frac{TP}{TP + FP}$$

- *Recall* is the sum of all the correctly marked cases divided by the sum of all the cases the system should have marked (TP+FN):

$$R = \frac{TP}{TP + FN}$$

- *Accuracy* is the sum of all correctly marked (TP) and the correctly unmarked cases (TN) divided by the sum of all cases under consideration:

$$A = \frac{TP + TN}{TP + FP + FN + TN}$$

- *Fmeasure* is the harmonic mean between *Precision* and *Recall*, according to each one an equal weight:

$$F = 2.\frac{P.R}{P + R}$$

Two types of evaluation can be performed with the results mentioned above: we can adopt either a *strict* or a *relaxed* evaluation scenario, depending on the weight accorded to the partial matches. In partial matches, three elements are at play: the connector tag, and the two clauses; while the tag is always correct, either one of the clauses might be wrong or incomplete. We have attributed an equal weight of 0.33 to each of the three elements. In the *relaxed* evaluation we considered the partially matched clauses to be correct whenever the intersection between the output and the reference is not null. In a *strict* evaluation, the partially matched clauses count as false positives (FP).

The nonexplicit connection between sentences, which links every two consecutive sentences that are not otherwise linked by any discourse connector, is represented by the symbol '&', as explained in Section 2. Initially, this relation was not taken into account for in the evaluation. This drastically decreases the values of Precision and Recall, as can be seen in Table 5, for most related sentences are not explicitly connected to previous discourse. The same applies to the relations between paragraphs (tagged with the symbol '&&').

In the evaluation *corpus*, a total of 108 neutral connections between sentences, and 65 relations between paragraphs, were tagged. All these relations were correctly tagged.

Adopting this evaluation strategy, the following results were obtained (Table 5). In the strict evaluation, results show a relatively good precision (82%) but a moderate recall (61%) yielding suboptimal values for accuracy (54%) a reasonable Fmeasure (70%). These values improve, but little, in the relaxed scenario, as only a small number of partial matches occurred (4).

Table 5: Results

	w/o & nor &&		**w/ &**		**w/ & and &&**	
	Strict	**Relaxed**	**Strict**	**Relaxed**	**Strict**	**Relaxed**
Precision (P)	0.82	0.90	0.94	0.97	0.96	0.98
Recall (R)	0.61	0.62	0.85	0.85	0.89	0.89
Accuracy (A)	0.54	0.58	0.81	0.82	0.86	0.87
Fmeasure (F)	0.70	0.73	0.89	0.90	0.92	0.93

The main reason for the high number of FN was the fact that some lexical items (15 instances), mostly compound subordinate conjunctions, were still missing in our lexicon[72]. Some of these lacunae are obvious and result only from the fact that these lexical items were dispersed in the STRING lexicons. Completing the lexicon will suffice to yield better results.

The second problem comes from the imbrication of clauses within the same sentence, as already mentioned in Section §2.4. The example below exemplifies this case:

> `[após, C0Conjsubordconsecutivetemporal` (*Assim, a fase de manutenção é a etapa do ciclo de vida do software na qual são efetuadas alterações no produto # sua liberação para o usuário.*)]

In this sentence, the parser first extracted the temporal subordinate conjunction *após* 'after' and, in the next step, it

[72] Namely, *ao passo que* 'but', *de fato* 'in fact', *de forma a* 'so that/in such a way that', *de maneira que* 'because', *de modo a* 'so that/in such a way that', *enquanto* 'while', *enquanto que* 'while', *quando* 'when', *porém* 'however'.

was not able to capture the causal conjunctive adverb *assim* 'thus'.

In our results, shown in Table 5, the impact of considering the '&' and '&&' operators between sentences and paragraphs, respectively, as been separately assessed. In the strict evaluation, the '&' connector produced an increase of 12% in Precision, 24% in Recall, 27% in Accuracy and 19% in Fmeasure. In the relaxed evaluation, results are only slightly better (from 1% to 3%) or nothing at all (Recall). This striking difference in the results shows the importance of this general cohesive process in the evaluation of discourse parsers.

On the other hand, and due to their smaller number, the fact of considering the default paragraph connector had but a small positive effect above the results already achieved with '&'. Compared with the results of the base evaluation (without & nor &&), in the strict scenario, one sees and increase of 14% in Precision, 28% in Recall, 32% for Accuracy, and 22% in Fmeasure. In a relaxed scenario, '&&' increases the base results in 8% Precision, 27% Recall, 28% Accuracy and 20% Fmeasure. This corresponds to just a slight improvement, from 1 to 5% against the results with only the operator &.

Similar works in the area (Pardo *et al.* 2004, Maziero *et al.* 2015) consider the nonexplicit connection between sentences in their results, without distinguishing the clause segmentation task from the mere sentence splitting general procedure and thus increasing the evaluation results.

As stated in Section 3, the coordinate conjunctions, such as the additive *e* 'and', were not accounted in this stage of the parser's development, given the difficulties in determining the coordination arguments and the clauses' boundaries, even in a fully POSdisambiguated text. To have a clearer idea of the problem, we considered parsing only the conjunction *e* 'and', including it in the targeted lexicon and applying it to the evaluation corpus. A total of 88 matches (out of 156) were tagged, and 72 of those correspond to cases of coordination of nominal or prepositional phrases', instead of clausal coordination. The following tagged sentence is an example of the coordination of subclausal constituents, instead of clauses:

> *A flexibilidade* **e** *a facilidade de uso de hiperdocumentos em_a Web têm garantido um futuro cada vez mais promissor para a utilização de sistemas de hipertexto.*
>
> **[E, C0Conjcoordadditive (***A flexibilidade* **#** *a facilidade de uso de hiperdocumentos em a Web têm garantido um futuro cada vez mais promissor para a utilização de sistemas de hipertexto***)]**

As can be seen, the word *e* 'and' was tagged as if it were introducing a clause, when it is actually linking two noun phrases. In STRING, this issue may be addressed, as coordination is so far treated at a strictly local level, between phrases. Hence, the coordination of clauses corresponds to a later stage of the interclause parsing.

A similar situation occurs due to POS classification of some discourse connectors, such as in the example below:

> **[após, C0Conjsubordtemporal (***Só é possível detectar esse tipo de erro* **#** *a análise da frase como um todo***.)]**

Certain words have different values, depending on what elements they introduce in a sentence. In this case, *após* 'after' is introducing a noun phrase, and therefore it should have been marked as a preposition. When this word introduces a clause, it has the value of conjunction. This is, in fact, the solution adopted in STRING. Because Unitex does not yet perform such disambiguation, the system considers it as a conjunction, thus producing a false positive (FP).

Considering this is still a prototype of a discourse parser, the results are quit satisfactory. This tool presents a fairly high Precision, both in the *strict* and in the *relaxed* evaluations. It still has a low Recall, but that can be improved by developing the tool further, particularly concerning issues found in embedded clauses and related, but independent sentences and paragraphs. Naturally the size of the evaluation sample is quite small and this results will have to be duplicated in a larger corpus and, eventually, from other text varieties.

6. CONCLUSIONS AND FUTURE WORK

This work presented a tool built for automatic discourse parsing, its main features and preliminary evaluation, which looks promising. This has proven to be a very difficult task, taking into account the existing POSambiguity in Portuguese, as shown in Section §4 and the effect of embedding of subclauses within sentences. Because of these difficulties, and in order to obtain a more accurate output, it is important to work with disambiguated text, where verbs are marked as being in the appropriate tense and other POS are also correctly tagged. This work on *corpus* annotation for lexicallyoriented, discursiverelated sentence relations, to the knowledge of the authors, has not been done to Portuguese yet.

One of the purposes of this paper was also to present the difficulty of the task at hand, and the challenges it poses for the task of content analysis. The relations addressed in this stage are relations between clauses within the same sentence or in adjacent sentences. In future developments of this tool more complexity will be added, by relating sentences and paragraphs in texts, improving the ability to analyse discourse.

Regarding future work, the prototype presented in this paper, when fully developed, may, on the one hand, contribute to the development of the STRING natural language processing chain; and, on the other hand, the automatic discourse parser can be improved by the use of the several modules of the STRING, as for example: lexical analysis (Vicente, 2013), identification of temporal expressions (Mauricio, 2011), and especially the anaphora resolution (Marques, 2013) and word sense disambiguation (Travanca, 2013; Suíssas 2014).

After further development, this tool may also be tested on other types of texts, with a less formal writing, to test its efficiency in other genres and text types. To sum up, a lot of work is yet to be done in the area of automatic discourse analysis, starting with automatic discourse segmentation. This paper is a modest contribution in that direction.

AKNOWLEDGMENTS

This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT), ref. UID/CEC/50021/2013.

REFERENCES

AitMokhtar, S., Chanod, J. and Roux, C. (2002). *Robustness Beyond Shallowness: Incremental Dependency Parsing*. Natural Language Engineering, 8(2/3): 121–144.

Berelson, B. (1952). *Content Analysis in Communication Research*. Glencoe: Free Press.

Cabrita, V. (2014). *Identificar, Ordenar e Relacionar Eventos*. Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

Clark, A., Fox, C., and Lappin S. (eds.), (2010). *The Handbook of Computational Linguistics and Natural Language Processing*, WileyBlackwell, Oxford.

Costa, J. (2009). *O Advérbio em Português Europeu*. Colibri. Lisboa.

Courtois, B. (1990). *Un Système de Dictionnaires Électroniques pour les Mots Simples du Français*. Langue française 87.

Diniz, C. (2010). *Um Conversor Baseado em Regras de Transformação Declarativas*. Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

Diniz, C., Mamede, N. and Pereira, J. (2010). *RuDriCo2 a Faster Disambiguator and Segmentation Modifier*. II Simpósio de Informática (INForum 2010): 573–584.

Dwight, H. (1948). *Power and Personality*. New York, NY.

Eleutério, S., Ranchhod, E., Freire, H. and Baptista, J. (1995). *A system of electronic dictionaries of Portuguese*. Lingvisticae Investigationes XIX: 1: 5782. John Benjamin B. V.. Amsterdam.

Fernandes, G. (2011). *Classification and Word Sense Disambiguation: The case of –mente ending adverbs in Brazilian Portuguese*. Master thesis, ErasmusMundus International Master on Natural Language Processing and Human Language Technologies, Universitat Autónoma de Barcelona/Universidade do Algarve.

Grice, P. (1975). *Logic and Conversation*. In Cole, P.; Morgan, J. L. (eds.). Syntax and Semantics 3: Speech Acts: 4158. Academic Press. New York.

Halliday, M. and Hasan, R. (1976). *Cohesion in English*. Essex: Longman.

Harris, Z. (1952). *Discourse Analysis*. Language 28: 130.

Harris, Z. (1991). *A Theory of Language and Information A Mathematical Approach*. Clarendon Press. Oxford.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology* (2nd ed.). Sage. Thousand Oaks, CA

Liu, B. (2012). *Sentiment analysis and opinion mining*. Synthesis lectures on human language technologies, 5(1), 1167.

Mamede, N., Baptista, J., Diniz, C. and Cabarrão, V. (2012). *STRING: A Hybrid Statistical and RuleBased Natural Language Processing Chain for Portuguese*. In Caseli, H., Villavicencio, A., Teixeira, A., and Perdigão, F., editors, Computational Processing of the Portuguese Language, Proceedings of the 10th International Conference, PROPOR 2012 Demo Sessions, volume Demo Session, Coimbra, Portugal.

Mann, W. and Thompson, S. (1988). *Rhetorical Structure Theory: Toward a Functional Theory of Text Organization*. Text 8 (3): 243281.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press. Cambridge, Massachusetts.

Marques, J. (2013). *Anaphora Resolution*. Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

Maurício, A. (2011). *Identificação, Classificação e Normalização de Expressões Temporais*. Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

Maziero, E., Hirst, G. and Pardo, T. (2015). *Adaptation of Discourse Parsing Models for the Portuguese Language*, 2015 Brazilian Conference on Intelligent Systems (BRACIS 2015), IEEE, pp. 140145. http://www.icmc.usp.br/~taspardo/BRACIS2015MazieroEtAl.pdf

Mendes, A. (2013). *Organização Textual e Articulação de Orações*, in Raposo *et al.* 2013: pp. 16911759.

Mitkov, R. (2014). *Anaphora Resolution*. Studies in Language and Linguistics. Taylor & Francis.

Molinier, C. and Levrier, F. (2000). *Grammaire des Adverbes: Description des Formes em 'ment'*. Droz. Genève.

Neuendorf, K. (2002). *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage.

Palma, C. (2009). *Estudo Contrastivo PortuguêsEspanhol de Expressões Fixas Adverbiais*. Master thesis, Universidade do Algarve. Faro, Portugal.

Pang, B., and Lee, L. (2008). *Opinion mining and sentiment analysis*. Foundations and trends in information retrieval, 2(12), 1135.

Pardo, T., Nunes, M. and Rino, L. (2004). *DiZer: An Automatic Discourse Analyzer for Brazilian Portuguese*. In the Proceedings of the 17th Brazilian Symposium on Artificial Intelligence – SBIA (Lecture Notes in Artificial Intelligence 3171): 224234. São LuisMA, Brazil.

Pardo, T. and Nunes, M. (2008). *On the Development and Evaluation of a Brazilian Portuguese Discourse Parser*. Revista de Informática Teórica e Aplicada, 15(2), 4364.

Paumier, S. (2003). *De la Reconnaissance de Formes Linguistiques a l'Analyse Syntaxique*. Volume 2, Manuel d'Unitex. Ph.D. thesis, IGM, Université de MarnelaVallée.

Paumier, S. (2016). *Unitex 3.1 User Manual*. Acceded in the 7th of March 2016, in: http://wwwigm.univmlv.fr/~unitex/

Ranchhod, E., Mota, E. and Baptista J. (1999). *A Computational Lexicon of Portuguese for Automatic Text Parsing*. In Proceedings of SIGLEX'99: Standardizing Lexical Resources, 37th Annual Meeting of the ACL: 7481, College Park, Maryland, USA.

Raposo, E., Nascimento, M., Mota, M., Segura, L. and Mendes, A (2013). *Gramática do Português*. Lisboa: Fundação Calouste Gulbenkian.

Resnik, P. and Lin, J. (2013). *The Handbook of Computational Linguistics and Natural Language Processing*. Chap. Evaluation of NLP Systems: 271–295. WileyBlackwell.

Sperber, D. and Wilson, D. (1986). *Relevance*. Oxford: Blackwell.

Suíssas, G. (2012). *MachineLearning Verb Sense Disambiguation*, Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

Tipaldo, G. (2013). *Handbook of TV Quality Assessment*. UCLan University Publishing. Preston, UK.

Tipaldo, G. (2014). *L'analisi del contenuto e i mass media*. Bologna, IT: Il Mulino.

Tofiloski, M., Brooke J. and Taboada, M. (2009). *A Syntactic and LexicalBased Discourse Segmenter*. In Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics. Singapore: 7780.

Travanca, T. (2013). *Disambiguation of Verb Senses*, Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

Vale, O. and Baptista, J. (2015). *Avaliação da flexão verbal do novo dicionário de formas flexionadas do UNITEXPB*. in: Claudia Freitas, Alexandre Rademaker (Eds.) STIL 2015, X Brazilian Symposium in Information and Human Language Technology and Collocated Events: 171180, Natal, Rio Grande do Norte, Brasil.

Vicente, A. (2013). *LexMan: um Segmentador e Analisador Morfológico com Transdutores*. Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.

Weber, R. (1990). *Basic Content Analysis*. 2nd ed.. Sage. Newbury Park, CA.

Wittgenstein, L. (1953). *Philosophical investigations*. Macmillan, New York.

Writing Center (2014). *Transitional Words and Phrases*, The Writer's Handbook. University of Wisconsin, Writing Center. Madison, Wisconsin: University of WisconsinMadison.