Ученые записки Крымского федерального университета имени В. И. Вернадского
Филологические науки. Том 2 (68). № 2. Ч. 1. 2016 г. С. 236–250.

UDC 811.134.3

TOWARDS DISCOURSE PARSING ON GRAMMATICAL PRINCIPLES

Talhadas R., Mamede N., Baptista J.

U. AlgarveFCHS/INESC ID Lisboa, rtalhadas@gmail.com
U. LisboaIST/INESC ID Lisboa, nuno.mamede@inescid.pt
U. AlgarveFCHS/INESC ID Lisboa, jbaptis@ualg.pt

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, Potuguese.

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 formalization, 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 tweets 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 in 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 of NLP 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; Mazieroet *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 at 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 (Tofiloskiet *al*., 2009) and this 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 linguistic issues raised in the construction of a discourse segmentation tool for European Portuguese. This is the first step towards the integration of such a tool into a fullyfledged, rulebased and statistical NLP system for Portuguese.

This paper is structured as follows: First, in Section 1, we present the main linguistic processes and lexical devices involved in the structuring of discourse by way of the socalled connectors. Then, in Section 2, 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. 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.

1. LINGUISTIC DEVICES 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). We will use as testing ground 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. This corpus has already undergone linguistic notation regarding its discursive structure, within a RST theoretical framework, though this data is not publicly available. Still, using this corpus as a workbench, we expect to be able, in the future, to comparedifferent analysis and theoretical approaches.

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 processing towards a transsentential, discourse parsing. The grammatical approach, here envisioned, tries to capture discourse structure in a ‘close-to-the-text’ (‘shallow’, if one wants to call it) manner, naturally flowing from a basic parsing stage; with limited theoretical constraints, to promote flexibility and reusability; and with limited semantic interpretation added, in order to maximize reproducibility.

In this paper, we focus on the use of two major types of connective devices: conjunctions (§1.1) and conjunctive adverbs (§1.2). 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 not, by all 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, 2013), 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.

1.1. 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 can be 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 (conjunctions linking phrases – and not clauses, are ignored at this stage). 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). Connective devices, mainly conjunctions, can relate clauses. 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.

1.2. 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 adequate and sufficient; with only 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 could help to organize semantically similar phenomena.

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

Another important aspect is that conjunctions, both simple and compound (i.e. multiword) constitute a finite 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 and phrases. To this paper, we used the (quite extensive) lexical 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**nãoconseguiuterminar antes dela**porque**ela é muitorápida.*

(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, and then by A and B), as shown in (2):

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

1.3. 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 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 often 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)
*O Pedro fez isto.A Ana, **porém**, (não) fez aquilo*
(Pedro did this. Ana, *however*, did (not_do) that)
O Pedro fez isto.A Ana (não) fez aquilo, ***porém***
(Pedro did this. Ana did (not_do) that, *however*)

Besides that, sentencemodifying adverbs cannot be extracted by clefting:

*A Ana fez aquilo, **porém*** (Ana did that, however)
Foiporém**quea Ana fez aquilo* (It was however that Ana did that)

In fact, this is an operation that can only be used to front sentenceinternal constituents:

*A Ana fez aquilo**hoje*** (Ana did that *today*)
*Foi**hoje**quea 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 context 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, poréma 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'.

1.4. 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 telefonouaofilho.*

(Pedro read the newspaper, watched tv for a while and phoned his sun)

(2) *O Pedro foi logo comprar um jornal. Hátrêsdiasquenãosabia 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 relationa type ofELABORATION (Pardo*et al.*, 2004).

1.5. Sentence sequences, clause embedding and ordering

Finally, it is relevant to mention some issues on sequences of sentences and clauses and the challenges this poses to an adequate representation of discourse structure.

Adverbial subordinate clauses can often be fronted to the beginning of the main clause of a sentence. Coordination, on the other hand, does not allow the permutation of coordinated clauses. Conjunctive adverbs usually link two consecutive sentences and appear at the beginning of the second one, but they have high mobility within the sentence in which they are. Finally, sequences of sentences without any explicit connector must still be related (by '&' or '&&'), so that all sentences in a discourse may be linked.

Furthermore, finding the boundaries of the argument clauses of both coordinate and subordinate clauses is not obvious, not only because this relies on a good parsing tools, but also, for the often ambiguous, non-local dependencies connective elements (especially coordinate conjunctions) may establish within the sentence.

Thus, complex sentences involving both parataxis and hypotaxis can give rise to complex combinatorial patterns. Even if an arbitrary limit is imposed on the sentence structure (and one fails to see a valid reason why it should be so, outside practical considerations), building a parser for such complex, often very imbricate, clausal structures is not a trivial task.

Hence, considering the example already provided in (2), the structure between clauses can be formalized as in (S), where the specific content of sentences is represented by S_i, leaving the connectors (conjunctions) in place, as follows:

(S) $[S_1 \textit{enquanto} S_2]_A \textit{mas}\ [S_3 \textit{porque}\ S_4]_B$

(One can also consider that S_i is not exactly the sentence but the topmost node of the sentence parsed so far, in a bottom–up approach).

Alternatively, the operators (conjunctions) may extracted from the sentence arguments, as in (P):

(P) $\textit{mas}\{[\textit{enquanto}\ (S_1\ , S_2)]_A\ ,[\textit{porque}\ (S_3\ , S_4)]_B\ \}$

This later solution (that we used for this paper) is more easily convertible into a graph-like structure (as in Bick 2000), where words would be the nodes and the dependencies the arcs between those nodes. This would allow for a graphical representation (see Harris 1991, for several complex examples), which could help human annotation of corpus to build (and evaluate) the discourse parser.

Any of these textual modifications, from the initial discourse (1) to its representations in (2), (S) or (P) 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»".

This, perhaps too long, discussion around the issues of formalisation is not at all some idle talk. One must bear in mind that not only this analysis must be entirely reproducible; it must also be humanly intelligible, even in very complex cases, with multiple combinations of hypotaxis and parataxis, in order to build the linguistic resources required to develop and evaluate such discourse parsers.

Whatever the formalism do be adopted (and the annotation tools to be developed), 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.

2. LINGUISTIC RESOURCES AND NLP TOOLS FOR PORTUGUESE

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

2.1. Linguistic resources

The lexicons of conjunctions and conjunctive adverbs of the STRING natural language processing chain (Mamede*et al.*, 2012) were adapted to the Dela-Unitex formalism, in order to use them with the linguistic development platform Unitex (Paumier, 2003, 2016). The choice of this system has to do with its simplicity and the fact that it was relatively easy to adapt current linguistic resources from STRING to be used with this system.

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*, *porestarazão*, *vistoisto*). 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:

```
afim de,.CONJ+subordinate+final
antesque,.CONJ+subordinate+temporal+anterior
depois de,.CONJ+subordinate+temporal+posterior
enquanto,.CONJ+subordinate+temporal+simultaneous
paraque,.CONJ+subordinate+final
porque,.CONJ+subordinate+causal
porcausa 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:

```
asaber,.ADV+Advconj+appositive
afinal de contas,.ADV+Advconj+consecutive
aindaassim,.ADV+Advconj+concessive
aindaporcima,.ADV+Advconj+additive
antes de mais,.ADV+Advconj+temporal
```

assim\,,.ADV+Advconj+causal
casocontrário,.ADV+Advconj+conditional
deresto,.ADV+Advconj+concessive
em o entanto,.ADV+Advconj+adversative
isto é,.ADV+Advconj+appositive
ouseja,.ADV+Advconj+appositive
porconseguinte,.ADV+Advconj+consecutive
porenquanto,.ADV+Advconj+temporal
porosvistos,.ADV+Advconj+causal
portanto,.ADV+Advconj+causal
querdizer,.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. 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.

2.2. Corpus

For the development of the parser, we intend to use the previously mentioned TCC *corpus* (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). This counting was 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.

This corpus was used at this stage as a source of the main types of discourse relations, since it has already been annotated for discursive relations among sentences and clauses, even if from a different theoretical perspective, in view of future comparison.

The *corpus* was pre-processed and the texts were split 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 (notice that colon <:> and semicolon <;> were not treated as sentence boundaries). 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.

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.

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

AdvConj	**Count**	**Conj**	**Count**
porexemplo	28	devido a	22
ouseja	10	paraque	21
em_oentanto	9	além de	19
por outro lado	8	quanto	17
assim,	7	bemcomo	7
portanto	6	umavezque	5
emseguida	3	nem	4
isto é	3	e/ou	5
porsuavez	2	apesar de	5
por um lado	2	embora	4

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.

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****maistarde um erro for encontradoem_oprocesso de desenvolvimento de software, maior é o custoparacorreção de esseerro.*

[quanto, C0Conjsubordcomparative (*Segundo Pressman, #maistarde um erro for encontrado em o processo de desenvolvimento de software, maior é o custoparacorreção de esseerro***.)]**

In this sentence,*quanto* is part of the proportional (discontinuous) conjunction *quantomais X, mais Y*. Because the program failed to identifythis 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.

CONCLUSIONS AND FUTURE WORK

This work presented the main linguistic issues and requirements towards building an automatic discourse parser with reference to Portuguese. This has proven to be a very difficult task, taking into account the existing POSambiguity in Portuguese and the effect of reordering and embedding of subclauses within sentences, the conjunctive adverb mobility within sentences, and the consideration of default connectors '&' between two, immediately sequential, but otherwise formally unrelated, sentences (or paragraphs). 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 more complexity must still be added, by relating sentences and paragraphs in texts, improving the ability to analyse discourse.

Regarding **future work**, we aim at the development of an automatic discourse parser integrated in the STRING natural language processing chain, using its rule-based parser XIP by way of dependency extraction rules, whose results approximate a graph-like representation. After development, this tool may also be tested on other types of texts, with a less formal writing, to test its efficiency and portability 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 Fundaçãopara a Ciência e a Tecnologia (FCT), ref. UID/CEC/50021/2013.

REFERENCES

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

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

3. Bick, Eckard. The Parsing System “PALAVRAS”. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Arhus University Press. (2000)

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

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

6. Costa, J..O Advérbio em PortuguêsEuropeu. Colibri. Lisboa. (2009)

7. Courtois, B..Un Système de DictionnairesÉlectroniques pour les Mots Simples du Français. Langue française 87. (1990)

8. Diniz, C..Um ConversorBaseado em Regras de TransformaçãoDeclarativas. Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.(2010)

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

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

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

12. Fernandes, G..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, UniversitatAutónoma de Barcelona/Universidade do Algarve. (2011)

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

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

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

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

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

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

19. Mamede, N., Baptista, J., Diniz, C. and Cabarrão, V..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. (2012)

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

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

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

23. Maziero, E., Hirst, G. and Pardo, T..Adaptation of Discourse Parsing Models for the Portuguese Language, 2015 Brazilian Conference on Intelligent Systems (BRACIS

2015), IEEE, pp. 140145.(2015)http://www.icmc.usp.br/~taspardo/BRACIS2015Maziero-EtAl.pdf

24. Mendes, A..Organização Textual e Articulação de Orações, in Raposoet al. 2013: pp. 16911759. (2013)

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

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

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

28. Palma, C..EstudoContrastivoPortuguêsEspanhol de ExpressõesFixasAdverbiais. Master thesis, Universidade do Algarve. Faro, Portugal. (2009)

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

30. Pardo, T.,Nunes, M. and Rino, L..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. (2004)

31. Pardo, T. and Nunes, M..On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. Revista de InformáticaTeórica e Aplicada, 15(2), 4364. (2008) http://www.icmc.usp.br/pessoas/taspardo/CorpusTCC.zip [20160330]

32. Paumier, S..De la Reconnaissance de FormesLinguistiquesal'AnalyseSyntaxique. Volume 2, Manuel d'Unitex. Ph.D. thesis, IGM, Université de MarnelaVallée. (2003)

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

34. Ranchhod, E.,Mota, E. and Baptista J..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. (1999)

35. Raposo, E., Nascimento, M., Mota, M., Segura, L. and Mendes, A. Gramática do Português. Lisboa: FundaçãoCalousteGulbenkian. (2013)

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

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

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

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

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

41. Vicente, A..LexMan: um Segmentador e AnalisadorMorfológicocomTransdutores.Master thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal. (2013)

42. Weber, R..Basic Content Analysis. 2nd ed.. Sage. Newbury Park, CA.

43. Wittgenstein, L. (1953). Philosophical investigations. Macmillan, New York. (1990)

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

К ВОПРОСУ О СИНТАКСИЧЕСКОМ АНАЛИЗЕ ДИСКУРСА НА ОСНОВЕ ГРАММАТИЧЕСКИХ ПРИНЦИПОВ

Талядаш Р., Мамеде Н., Батишта Ж.

Университет Альгарве, факультет социальных и гуманитарных наук, Фару/ Институт инженерии и компьютерных систем, Лиссабон, Португалия, rtalhadas@gmail.com
Лиссабонский университет – Институт технических наук / Институт инженерии и компьютерных систем, Лиссабон, Португалия, nuno.mamede@inescid.pt
Университет Альгарве, факультет социальных и гуманитарных наук, Фару/ Институт инженерии и компьютерных систем, Лиссабон, Португалия, jbaptis@ualg.pt

Контент-анализ является важным методом анализа для многих гуманитарных и социальных наук, включая психологию и социологию. Выявление структур текста является значимым шагом в определении, как большинство содержательных элементов организовано. Кроме сегментации текста на абзацы, предложения, придаточные предложения, использование связующих элементов дискурса является фундаментальным элементом для структурирования текста. Эти связующие звенья включают союзы и союзные наречия, и они выявляют значимые отношения между предложениями, которые образуют текст. В этой статье мы иллюстрируем применение этого метода для определения важных компонентов текста и их подробную организацию. Для оценки применяется метод синтаксического анализа, но он также может быть применен для выполнения многих задач при контент-анализе. Междисциплинарный метод связывает темы лингвистики и компьютерной лингвистики с возможным применением в нескольких сферах социальных наук, в которых контент-анализ и синтаксический анализ могут быть важными.

Ключевые слова: контент-анализ, текстологический / синтаксический анализ, связующие элементы дискурса, португальский язык.