DOI: 10.14393/DL52-v16n4a2022-10

Evaluating a typology of signals for automatic detection of complementarity

Avaliação de uma tipologia de sinais para a detecção automática da complementaridade

Jackson Wilke da Cruz SOUZA[*]
Ariani DI FELIPPO[**]

ABSTRACT: In a cluster of news texts on the same event, two sentences from different documents might express different multi-document phenomena (redundancy, complementarity, and contradiction). Cross-Document Structure Theory (CST) provides labels to explicitly represent these phenomena. The automatic identification of the multi-document phenomena and their correspondent CST relations is definitely handy for Automatic Multi-Document Summarization since it helps computers understand text meaning. In this paper, we evaluated a typology of (textual) signals for the automatic detection of the CST relations of complementarity (i.e., *Historical background*, *Follow-up* and *Elaboration*) in a multi-document corpus of news texts in Brazilian Portuguese. Using algorithms from different machine-learning paradigms, we obtained classifiers that achieved high general accuracy (higher than 90%), indicating the potential of the signals.

RESUMO: Em uma coleção de notícias sobre um mesmo evento, duas sentenças de textos distintos podem expressar diferentes fenômenos multidocumento (redundância, complementaridade e contradição). A *Cross-Document Structure Theory* (CST) provê rótulos para representar esses fenômenos. A identificação automática dos fenômenos multidocumento e das relações CST correspondentes é central à Sumarização Automática Multidocumento, pois ajuda a máquina a entender o conteúdo textual. Neste artigo, avaliou-se uma tipologia de sinais (textuais) para a detecção automática das relações CST de complementaridade (*Historical background*, *Follow-up* e *Elaboration*) em um *corpus* multidocumento de notícias em Português do Brasil. Utilizando algoritmos de diferentes paradigmas de Aprendizado de Máquina, obtiveram-se classificadores que atingiram alto índice de acurácia geral (superior a 90%), indicando o potencial dos sinais.

[*] PhD in Linguistics (UFSCar), professor in Instituto de Ciências Sociais Aplicadas at Universidade Federal de Alfenas (UNIFAL-MG). ORCID: https://orcid.org/0000-0003-1881-6780. jackcruzsouza@gmail.com

[**] PhD in Linguistics (UNESP), professor in Departamento de Letras at Universidade Federal de São Carlos (UFSCar). ORCID: https://orcid.org/0000-0002-4566-9352. arianidf@gmail.com

KEYWORDS: *Cross-Document Structure Theory.* Automatic summarization. Multi-document Corpus. Complementarity. Textual signal.

PALAVRAS-CHAVE: *Cross-Document Structure Theory.* Sumarização automática. Complementaridade. *Corpus* multidocumento. Sinal textual.

1 Introduction

Since the estimated size of the indexed Web is around 3.7 billion pages[1], its amount of textual information has already exceeded human limits of manageability. Given this scenario, subareas of Natural Language Processing (NLP) can produce computational solutions to deal with this large amount of data available to the user. Sub-areas dealing with content production and selection are those that have gained prominence in recent years, producing, for example, sentiment analysis, question and answer systems and automatic summarizers.

Specifically Automatic Multi-Document Summarization (MDS) NLP application that may assist users in acquiring relevant information in a short time. MDS aims at identifying the main information in a cluster of texts and presenting it as a summary (MANI, 2001). Much of the work to date has focused on extracts, i.e., summaries produced by concatenating sentences taken exactly as they appear in the source-documents (NENKOVA; MACKEOWN, 2011).

One important theory that guides extractive methods is Cross-Document Structure Theory (CST) (RADEV, 2000). It proposes relations to connect sentences from topically related texts. Such relations can be grouped into two categories (MAZIERO; JORGE; PARDO, 2010). Content relation indicates similarities and differences between sentences (*Identity*, *Equivalence*, *Summary*, *Overlap*, and *Subsumption*), complementarity (*Historical background*, *Follow-up*, and *Elaboration*), and contradiction (*Contradiction*). The form category conveys relations (*Indirect-speech*, *Modality*, *Attribution*, *Citation*, and *Translation*) that deal with shallow aspects of texts.

[1] Available from: https://www.worldwidewebsize.com. Access in 01/09/2021.

In MDS, there are very important challenges such as capturing the most important information of a topic within a generic perspective or prioritizing information preferences specified by the user (such as context information or the evolution of an event in time). CST annotation provides the means to deal with that, since the relations allow for detecting the multi-document phenomena (redundancy, complementarity and contradiction). Relations as Equivalence and Identity, for example, help to exclude repeated information, since a coherent extract should not have redundancy. Otherwise, if a user requires more context information about an event in the summary, the Historical background relation is helpful.

There have been many efforts to automatically detect the CST relations (e.g., ZHANG; BLAIR-GOLDENSHON; RADEV, 2002; ZHANG; OTTERBACHER; RADEV, 2003; MAYABE; TAKAMURA; OKUMURA, 2008; ZAHRI;, FUKUMOTO, 2011; KUMAR; SALIM; RAZA, 2012). One of them is CSTParser (MAZIERO; PARDO, 2012), an online multi-document parser based on CST for Brazilian Portuguese (BP). Using machine learning (ML) techniques, the system detects the relations with a general accuracy of 68,13%. Except for *Contradiction* and *Identity* (which are detected by rules), the parser uses similarity features to decide which CST relation (including those of complementarity) is held between sentences, since this type of relation only occurs between semantically related sentences.

Since complementary content might be important to build a multi-document extract and similarity is not sufficient for recognizing the different types (or CST relations) of complementarity, some efforts have been made in the last years to provide more extensive descriptions about the phenomenon and more efficient automatic methods for detecting it (SOUZA; DI-FELIPPO, 2018; SOUZA, 2015, 2019, 2021).

In Souza (2019), a corpus annotation of temporal markers and a wide variety of other (textual) signals of complementarity was carried out that resulted in a typology. It contributes to better understanding how complementarity is marked in the text, and

also provides attributes that can be used by automatic classifiers to recognize the different types (and CST relations) of complementarity. Recently, Souza (2021) refined the typology by exploring other aspects that seem to guide the readers to recognize complementary relations that hold between sentences.

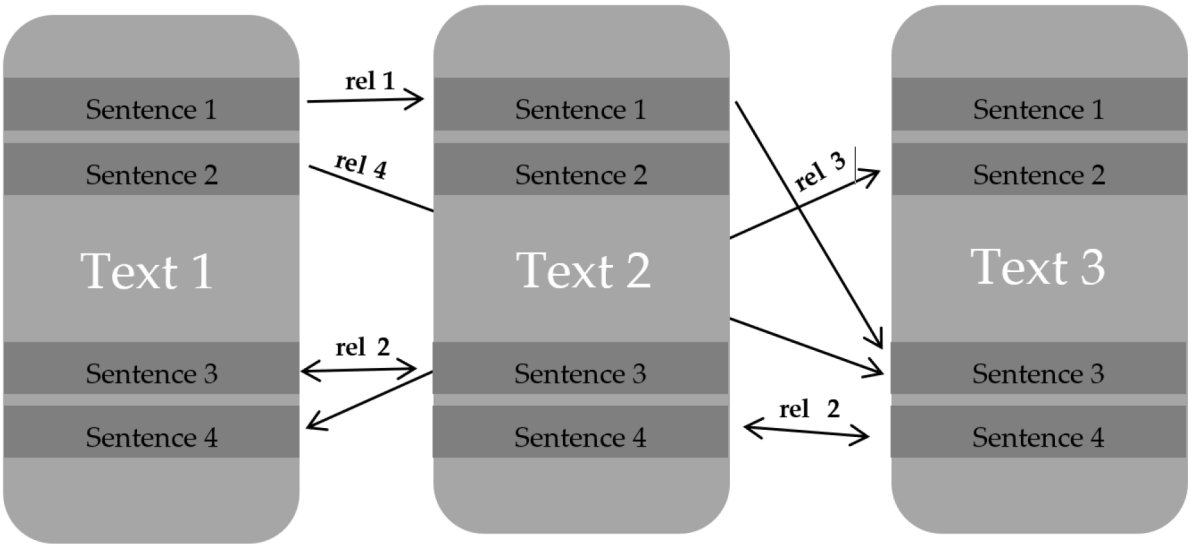
In this paper, we focused on evaluating the typology of Souza (2019) for the automatic detection of the CST relations of complementarity (i.e., *Historical background, Follow-up* and *Elaboration*) in a multi-document corpus of news in BP. The evaluation was performed using algorithms from different ML paradigms, and the results are promising.

Following the Introduction, Section 2 provides a brief introduction to CST, the main framework for the analysis, and an overview of the notion of complementarity. Section 3 presents the CSTNews corpus and the typology of signals. Section 4 investigates, through ML algorithms, the use of signals to distinguish the CST relations of complementarity. Finally, Section 5 summarizes the paper, and presents a few directions for future research.

2 Related works

Two sentences from topically related texts can be similar and different in several ways. One of the most relevant models to represent multi-document relations is CST, which was inspired by Rhetorical Structure Theory (RST) (MANN; THOMPSON, 1987). The difference between theories is that RST is aimed at capturing the rhetorical relation between adjacent text units while CST goes across topically related texts. Figure 1 illustrates a generic multi-document analysis at the sentence level.

Figure 1 – Generic scheme of multi-document analysis.



Source: Maziero (2012).

The original version of CST includes a set of 24 relations (RADEV, 2000) (Table 1).

Table 1 – Original set of CST relations.

Identity	Modality	Judgment
Equivalence	Attribution	Fulfillment
Translation	Summary	Description
Subsumption	Follow-up	Reader profile
Contradiction	Elaboration	Contrast
Historical background	Indirect speech	Parallel
Cross-reference	Refinement	Generalization
Citation	Agreement	Change of perspective

Source: based on Radev (2000).

In the last decade, interest in CST applications began to arise, especially in MDS, but also in other areas such as Query Reformulation, Learning Support, and Opinion Mining in the web (e.g., BELTRAME; CURY; MENEZES, 2012, INAM *et al.*, 2012; MURAKAMI *et al.*, 2010).

In order to obtain a better formalization and improving annotation concordance (by reducing ambiguity), Maziero, Jorge and Pardo (2010) proposed, based on the annotation of CSTNews, a typology for 14 CST relations according to their semantic

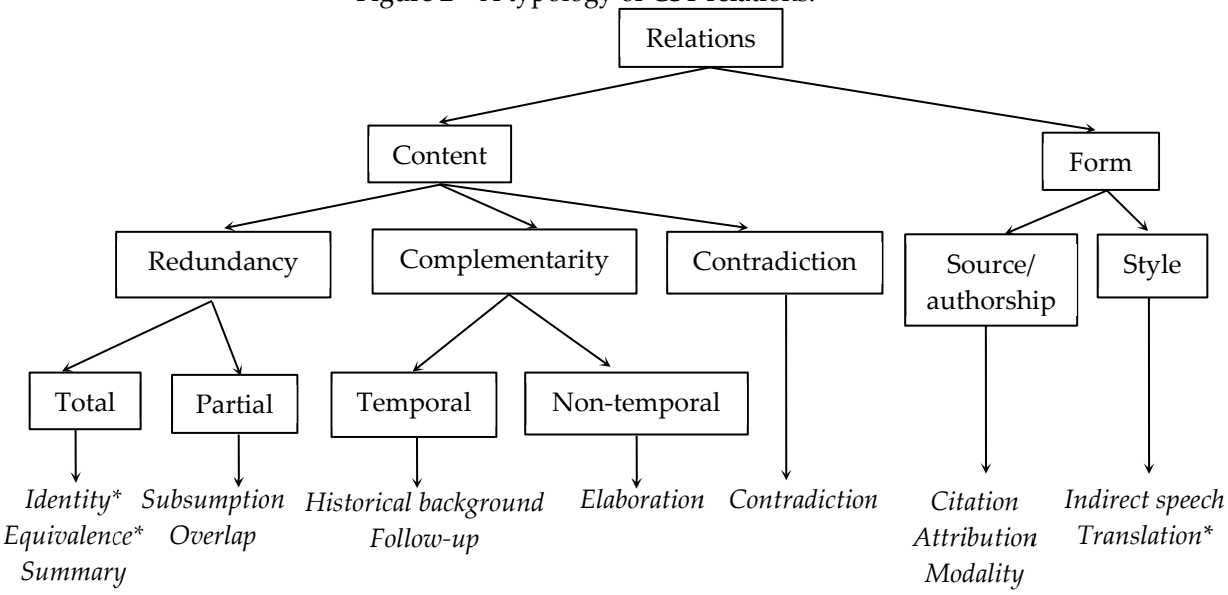
nature. This typology is illustrated in Figure 2, where the CST relations are at the lowest level of the hierarchy. One may observe in Figure 2 that there are two main categories of relations in the typology: content and form.

The content category refers to relations that indicate or capture (informational) similarities and differences among sentences. This category is divided into 3 subcategories: redundancy, complementarity, and contradiction. Redundancy includes relations that express total (i.e., *Identity*, *Equivalence* and *Summary*) or partial (i.e., *Overlap* and *Subsumption*) similarity among sentences. Complementarity relations link segments that elaborate, give continuity or background to some other information. *Historical background* and *Follow-up* are considered temporal, while *Elaboration* is non-temporal. The last subcategory only includes *Contradiction*.

The form category includes all the relations that deal with superficial aspects of information, for example, writing styles (*Indirect-speech*, *Modality*), citations (*Attribution*, *Citation*) or language (*Translation*). CST relations may also have directionality, being classified as symmetric or non-symmetric. In Figure 2, the asterisk indicates the symmetric relations, since one may read them in any direction.

To illustrate how complementarity relations occur among texts, Table 2 shows examples extracted from CSTNews. Each pair in Table 2 was selected from distinct news on the same event (i.e. from distinct news of the same cluster). The pair of sentences (S1 and S2) in (1) illustrates *Historical background*. In this case, the sentences were extracted from cluster C1, which comprises news reporting "a plane crash in Congo". S1 informs the place and the number of victims, while S2 provides a historical setting about S1 (i.e., air accidents are routine in the history of Congo).

Figure 2 – A typology of CST relations.



Source: Maziero, Jorge and Pardo (2010).

Table 2 – Example of types and CST relations of complementary from CSTNews[2].

Type	Relation	Pair of sentences
Temporal	Historical background	(1) S1: A plane crash in Bukavu, in the Eastern Democratic Republic of the Congo, killed 17 people on Thursday, said a United Nations spokesman on Friday. *(Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira, informou nesta sexta-feira um porta-voz das Nações Unidas)* S2: Air accidents are frequent in Congo, where 51 private companies operate elderly planes built in the former Soviet Union. *(Acidentes aéreos são frequentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética)*
	Follow-up	(2) S1: In his speech, Lula put emphasis on the end of agricultural protectionism. *(O discurso de Lula na ONU deu grande ênfase ao fim do protecionismo agrícola.)* S2: After Lula, it was the turn of the U.S. President George W. Bush to address the United Nations 62nd General Assembly. *(Depois de Lula, foi a vez do presidente americano George W. Bush discursar na Assembleia Geral da ONU)*

[2] In this paper, we first present an English translated version of the example, followed by the original one in Portuguese language.

Non-Temporal	Elaboration	(3) S1: The victims of the accident were 14 passengers and three crew members. *(As vítimas do acidente foram 14 passageiros e três membros da tripulação.)* S2: According to air traffic control, all crew members were Russian nationals. *(Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.)*

Example (2) illustrates the *Follow-up* relation, since S2 presents additional information which has happened since S1. The sentences were compiled from news on "the speech of the former Brazilian president, Luiz Inácio Lula da Silva, at the 62nd session of the United Nations General Assembly". In this case, *Follow-up* is signaled by the temporal expression "after Lula" in S2.

The sentences in example (3) were also compiled from the cluster C1, and they illustrate *Elaboration*. More specifically, S1 presents the number and profile of the victims, while S2 details an element present in S1 without redundancy. In this case, S2 provides the nationality of the crew members who died in the crash.

Given its usefulness, especially for MDS applications, some efforts have been made in the last years to provide extensive descriptions about the multi-document phenomena covered by the CST relations. The first work on MDS in PB, by Maziero, Jorge and Pardo (2014), addressed the automatic detection of CST relations, including those of complementarity. Following the literature, the authors only used features that track some form of repetition or redundancy (such as the difference in number of words, percentage of words in common, number of words in the longest common substring, difference in the number of nouns, etc.) to predict the relations between two sentences. The ML techniques explored for the prediction of the content relations achieved a general accuracy of 70,51%, which is considered a good result given the subjectivity of the CST analysis.

With the common goal of automatically identifying and characterizing CST relations in texts, a few computational-linguistic studies have explored a wide array of linguistic or textual features of complementarity (SOUZA; DI-FELIPPO, 2018; SOUZA,

2015, 2019, 2021). More specifically, Souza (2015) and Souza and Di-Felippo (2018) developed the first corpus description of complementarity using 90 sentence pairs compiled from the CSTNews corpus. In this work, a set of 7 potential attributes was explored for temporal complementarity detection. The attributes are: noun overlap, sentence distance[3], subtopic overlap, temporal expression in S1, temporal expression in S2, adverb in S1, and adverb in S2. With the exception of temporal expressions and adverbs, the majority of the attributes is based on lexical similarity because it is known that the CST relations only occur between semantically related sentences.

Using some ML algorithms from Weka (Waikato Environment for Knowledge Analysis) (WITTEN; FRANK 2005), the potential of the attributes to discriminate between *Historical background* and *Follow-up* were evaluated. The JRip classifier learned the smallest set of rules with the highest general accuracy (80%). Among the 5 rules of JRip, three of them are based on "temporal expression in S2" to classify the *Follow-up* pairs. For attribute selection[4], the InfoGainAttributeEval algorithm was applied, and it also indicated the relevance of this feature (i.e., temporal expression in S2) in the task.

Later, Souza (2019) expanded the previous work by investigating temporal and non-temporal complementarities and a wide variety of signals (morphological, syntactic, semantic, and pragmatic). As a result, the signals were organized into a typology, yielding a hierarchical structure of textual cues.

More recently, in a study with no computational motivations, Souza (2021) refined the typology by exploring other aspects (not expressed in the text) on how the

[3] The relative distance between complementary sentences according to the position of the sentences in their correspondent source text. For example, given a sentence pair from a cluster x (S1 and S2), where S1 is Sentence 6 from Text 1 and S2 is Sentence 4 from Text 2, the distance value between them is equal to 2 (positions). The authors normalized this value by dividing it by the longest distance between two sentences identified in the subcorpus of complementarity relations.

[4] Attribute selection aims at improving the performance of the ML algorithms by removing irrelevant attributes, which reduces the processing time and generates simpler models.

complementarity between sentences is recognized by readers, such as the reading of their source texts.

In the next section, we briefly present the CSTNews corpus. We also present the mentioned typology (or taxonomy) of signals in detail, as it is the focus of the evaluation.

3 The CSTNews corpus and the typology of signals of complementarity

The study of Souza (2019) was conducted over CSTNews, a multi-document corpus contained 50 clusters of news in BP, totaling 140 texts, 2,088 sentences and 47,240 words (CARDOSO *et al.* 2011). The clusters are organized into 6 categories: world, politics, daily news, science, money, and sports. The source texts were compiled from the online news agencies *Folha de São Paulo*, *Estadão*, *O Globo*, *Jornal do Brasil*, and *Gazeta do Povo*.

Each cluster contains: (i) 2 or 3 texts; (ii) mono-document abstracts (produced by a human summarizer); (iii) 12 manual multi-document summaries (6 abstracts and 6 extracts); (iv) 1 automatic multi-document extract, and several types of annotation, corresponding to different levels of linguistic analysis.

One of the annotations is the manual identification of the CST relations across the input documents within the clusters. This annotation was performed by 4 computational linguists using the CSTTool editor (ALEIXO; PARDO, 2008) and the typology shown in Figure 2. From the total of 1,641 sentence pairs annotated in CSTNews, there are 713 pairs of complementarity, distributed in 370 of temporal (i.e., 77 of *Historical Background* and 293 of *Follow-up*), and 343 of non-temporal relations (i.e., *Elaboration*). Thus, complementarity corresponds to 43.44% of the total relations in the corpus.

For the study of (temporal and non-temporal) complementarity, Souza (2019) only used 655 pairs (i.e., 76 of *Historical Background*, 260 of *Follow-up* and 319 of

Elaboration), since he excluded 58 pairs for disagreeing with the original annotation. The main goal of this work was to explore how many, and what types of cues could be found if the signaling information was studied beyond temporal expressions and lexical similarity. The most important aspect of the description was to select and classify the types of cues.

For the description of these relations, a sequence of three main tasks was performed: (i) analysis of each sentence pair in the subcorpus of complementarity from CSTNews, (ii) delimitation (with brackets) of the signals involved in complementarity to indicate the specific CST relation, and, finally, (iii) documentation on how the relation is signaled.

We illustrate these tasks with a *Historical background* pair from the CSTNews corpus (4). This type of complementarity is codified by a CST relation with directionality (from S2 to S1), which means that S2 provides the contextual information about S1. According to the author´s analysis, the previous annotation of the sentence pair with *Historical background* is strongly related to the textual segment "foi o pior do país desde 1995" ("it was the country´s worst (earthquake)") (in parentheses) that occurs in S2. This segment contains 3 different textual signals (in bold), which are delimitated by square brackets and indexed according to the position of occurrence in the sentence: (i) superlative expression, (ii) preposition "desde", and (iii) named entity ("1995").

In the coding or delimitation task (ii), the author added signaling information to the existing relations from the CSTNews corpus. Then, the signals identified were extracted, and documented along with relevant information about their function. For example, the superlative expression ("foi o pior (do país)") directly refers back to the earthquake magnitude mentioned in S1, and provides a relative description about it, i.e., very high size or amplitude in comparison to the last strongest one. The remaining two signals, "since" and "1995", are used (together) to specify the particular time in

the past when the last strongest earthquake has happened. The author documented or described the signaling information in a separate Excel file.

A detailed description of the annotation for the sentence pair in (4) is provided in Table 3.

(4) S1: In the case of Japan, the mentioned magnitude of 6.8 is considered "strong".
("*No caso do Japão, a magnitude apontada de 6,8 é considerada "forte".*")

S2: ([**It was the country´s worst** (earthquake)]$_1$ **[since]**$_2$ [**1995**]$_3$), when a 7.3 magnitude earthquake killed more than 6,400 people in the city of Kobe.
(*"Foi o pior do país desde 1995, quando um tremor de magnitude 7,3 matou mais de 6.400 pessoas na cidade de Kobe.").*

Table 3 – Example of signals description.

Cluster/ Pair	Relation name	Signal type	Specific signal	Explanation – how the relation is signalled
32/52	Historical background	Main clause	Superlative	The expression "the worst" in main clause of S2 is used to compare the magnitude mentioned in S1 to the last strongest earthquake
		Word class	Preposition	"Since" is used to situate the event in time, specifically in relation to the year of the last strongest earthquake
		Time	Named entity	The NE "1995" functions as a signal because it refers back to a specific point in time.

After describing the 655 pairs, all signals were hierarchically organized in 3 levels (i.e., signal class, signal type and specific signal), following Taboada and Das (2013) and Das and Taboada (2018). The signal class is the top-level classification, and it has 5 tags representing the major classes of signals (i.e., referential, morphological, syntactic, semantic, and pragmatic), as described in Table 4. For each class, a second level is defined; for example, the class referential has only one type (anaphor), while

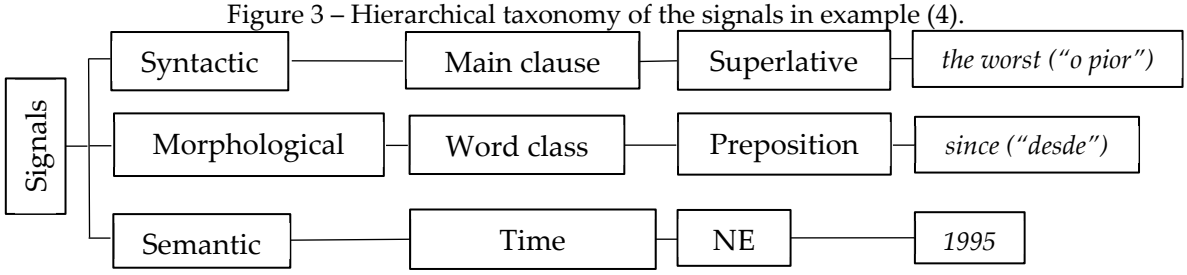
the semantic class is divided into 4 types (i.e., semantic field, semantic relation, addition, and time). Finally, the third level in the hierarchy refers to specific signals. The anaphoric type, for example, has two specific signals: associative and nominal.

Table 4 – Description of the different classes of complementarity signals[5].

Class	Description
Referential	Features include links where entities, similar or dissimilar, help interpret the relations.
Morphological	Among morphological features, tense is a very prominent feature, indicating temporal relations between the sentences.
Syntactic	At the syntactic level, there are a host of constructions that help identify relation. For example, subornative clauses tend to express details of some information given more generally in the other sentence of a pair.
Semantic	A semantic feature has two components, each belonging to one of the sentences. The components are in a semantic relationship with each other, such as *meronymy*, and *semantic field*.
Pragmatic	At the pragmatic level, there are several signals to guide the interpretation of relations. In the case of the news genre (which all the sentences/texts in the corpus belong to), the complementary information might be a list of items that details an event, a similar fact, a posterior or future event, etc.

The taxonomy of the signals in example (4) is provided in Figure 3.

Figure 3 – Hierarchical taxonomy of the signals in example (4).



We show the hierarchical organization of the signaling taxonomy in Table 4 as well as descriptive statistics of the frequency of signals in our corpus of

[5] For a more detailed description about the classes, types, and specific signals of the typology, along with example from the corpus, see Souza (2019, 2021).

complementarity. According to Table 5, the corpus includes 2,022 signals, distributed in 30 different specific signals. Referential and morphological classes are the less frequent, with a very similar distribution (i.e., 15.9% and 15.6%, respectively); pragmatic is the most frequent (25.4%), closely followed by syntactic (22.2%) and semantic classes (21%).

In Table 6, there is a summary of the relationship between CST relations and signals. The summary provides descriptive statistics of the frequency of the relations and how often each of them is signaled by specific cues.

We would like to point out that what Souza (2019) has found are positive signals, and this does not mean that the signals are used exclusively to indicate the relation. In other words, this means that relation identification, by humans and machines, relies or can rely on signals as indicators that a relation is present, but there are many-to-many correspondences between relations and signals, as we can see in Table 6. Although, the occurrence of certain signals seems to be typical of a relation. This is the case for the *superlative expression* in sentence 2 of a pair, which is a syntactic signal that only indicates *Historical background* (see example (4)). Also, the concept of a signaling information as an indicator of a relation also means that the signals, as textual devices, are not exclusively used to mark a CST relation; they may well have other purposes in the document or text. For instance, an associative anaphora, as a type of the referential class, also contributes to cohesion, in addition to signaling a relation.

In other words, we can say that textual signals are compatible with a CST relation, not necessarily indicators of the relation exclusively. However, the results seem to provide evidence that relation signaling is widespread and has potential for computational applications.

Table 5 – Typology of complementarity signals and its statistics.

Class	Signal type	Specific signal	Total		%	
Referential (322 – 15,9%)	Anaphor	Associative	81	322	4,0	15.9
		Nominal	241		11.9	
Morphological (315 – 15,6%)	Word class	Numeral	48	315	2.4	15.6
		Noun	17		0.8	
		Preposition	24		1.2	
		Verbal tense	149		7.4	
		Elocution verb	77		3.8	
Syntactic (450 – 22,2%)	Simple period (or main clause)	Adverbial phrase	73	99	3.6	4.9
		Superlative expression	26		1.3	
	Compound period	Reported speech	119	240	5.9	11.9
		Additional clause	28		1.4	
		Explanation clause	49		2.4	
		Direct object clause	29		1.4	
		Reduced relative clause	15		0.7	
	Displacement	Theme-Rheme	111	111	5.5	5.4
Semantic (422 – 20,9%)	Semantic field	Related words	63	63	3.1	3.1
	Semantic relation	Cause-effect	35	112	1.7	5,5
		Hyponymy (*is-a*)	20		1.0	
		Meronymy (*part-whole*)	57		2.8	
	Addition	Indicative word/phrase	170	170	8.4	8.4
	Time	Named entity (NE)	77	77	3.8	3.8
Pragmatic (495 – 25,4%)	Genre (about the event)	Detailing (*list of items*)	162	368	8.0	18,2
		Posteriority (*later event*)	92		4.5	
		Futurity (*future event*)	57		2.8	
		Continuity (*continuous event*)	18		0.9	
		Similarity	39		1.9	
	Argumentation	Focus	17	17	0.8	0.8
	Addition	Related information	52	52	2.6	2.6
	Aspectuality	Punctual event	38	76	1.9	3.8
		Durative event	38		1.9	

Source: Souza (2019).

Table 6 – Distribution of CST relations by signaling devices. Source: Souza (2019).

Tipology			**CST relations of complementarity**			**Total**
Class	**Signal type**	**Specific signal**	***Elab.***	***Follow-up***	***Historical background***	
Referential	Anaphor	Associative	50	31	0	81
		Nominal	132	89	20	241
Morph.	Word class	Numeral	11	35	2	48
		Noun	2	15	0	17
		Preposition	7	0	17	24
		Verbal tense	12	134	3	149
		Elocution verb	26	51	0	77
Syntactic	Simple period (or main clause)	Adverbial phrase	31	40	2	73
		Superlative expression	0	0	26	26
	Compound period	Reported speech	67	52	0	119
		Additional clause	26	2	0	28
		Explanation clause	37	5	7	49
		Direct object clause	22	7	0	29
		Reduced relative clause	12	3	0	15
	Displacement	Theme-Rheme	108	1	2	111
Semantic	Semantic field	Related words	29	34	0	63
	Semantic relation	Cause-effect	12	23	0	35
		Hyponymy (*is-a*)	16	4	0	20
		Meronymy (*part-whole*)	42	15	0	57
	Addition	Indicative word/phrase	4	109	57	170
	Time	Named entity (NE)	27	42	8	77
Pragmatic	Genre (about the event)	Detailing (*list of items*)	103	59	0	162
		Posteriority (*later event*)	0	92	0	92
		Futurity (*future event*)	0	57	0	57
		Continuity (*continuous event*)	0	18	0	18
		Similarity	0	0	39	39
	Argumentation	Focus	17	0	0	17
	Addition	Related information	52	0	0	52
	Aspectuality	Punctual event	0	0	38	38
		Durative event	17	0	0	17

Since Souza (2019) has proposed the typology or taxonomy of complementarity signals, we were interested in automatic detecting the CST relations of complementarity based on the signals. Thus, we investigated the discriminative power of the annotated signals.

4 Automatic validation of the signals

In order to determine whether certain signals (or combinations of them) predicate a CST relation of complementarity, we used ML algorithms available in Weka, which is a state-of-art facility for developing ML techniques and their application to real-world data mining tasks (WITTEN; FRANK, 2005). We conducted such automatic study because ML techniques consider every combination of signals to predict the classes (i.e., the CST relations of complementarity).

In this work, we used supervised ML algorithms, which basically map a function from known input-output pairs to estimate relationships between them. Most fundamentally, supervised learning utilizes a data set which includes both input features as well as the output class (or target) which are labeled at the start of training. In our case, each instance of the training data consisting of a sentence pair and its signals (i.e., the features) and the desired CST relation of complementarity (i.e., the output classes). Thus, the algorithms train on the input data set to produce a model which will differentiate among the output labels based on the most relevant attributes. In other words, the algorithms analyze the training data and produce a classifier that should be able to predict the correct classes (MITCHELL, 1997).

For using Weka, we converted the descriptive information illustrated in Table 3 into an ARFF (Attribute-Relation File Format) file, which is the most common format for data used in Weka (Figure 4). According to Figure 4, an ARFF file has two parts. The first one is a Header describing what each data instance should be like, and the second part is the Data (entry).

Figure 4 – Illustration of the ARFF file format for ML.

@ATTRIBUTE PAIR NUMERIC
@ATTRIBUTE RELATION {ELABORATION, FOLLOW_UP, HISTORICAL_BACKGROUND}
@ATTRIBUTE ASSOCIATIVE_ANAPHOR {YES, NO}
@ATTRIBUTE NOMINAL_ANAPHOR {YES, NO}
@ATTRIBUTE NUMERAL {YES, NO}
@ATTRIBUTE NOUN {YES, NO}
@ATTRIBUTE PREPOSITION {YES, NO}
@ATTRIBUTE VERBAL_TENSE {YES, NO}
@ATTRIBUTE ELOCUTION_VERB {YES, NO}
@ATTRIBUTE ADVERBAL_PHRASE {YES, NO}
@ATTRIBUTE SUPERLATIVE {YES, NO}
@ATTRIBUTE REPORTED_SPEECH {YES, NO}
@ATTRIBUTE ADDITIONAL_CLAUSE {YES, NO}
@ATTRIBUTE EXPLANATION_CLAUSE {YES, NO}
@ATTRIBUTE DIRECT_OBJECT_CLAUSE {YES, NO}
@ATTRIBUTE REDUCED_RELATIVE_CLAUSE {YES, NO}
@ATTRIBUTE THEME-RHEME {YES, NO}
@ATTRIBUTE RELATED_WORDS {YES, NO}
@ATTRIBUTE CAUSE-EFFECT {YES, NO}
@ATTRIBUTE HYPONYMY {YES, NO}
@ATTRIBUTE MERONYMY {YES, NO}
@ATTRIBUTE INDICATIVE_WORD {YES, NO}
@ATTRIBUTE NAMED_ENTITY {YES, NO}
@ATTRIBUTE DETAILING {YES, NO}
@ATTRIBUTE POSTERIORITY {YES, NO}
@ATTRIBUTE FUTURITY {YES, NO}
@ATTRIBUTE CONTINUITY {YES, NO}
@ATTRIBUTE SIMILARITY {YES, NO}
@ATTRIBUTE FOCUS {YES, NO}
@ATTRIBUTE RELATED_INFORMATION {YES, NO}
@ATTRIBUTE PUNCTUAL_EVENT {YES, NO}
@ATTRIBUTE DURATIVE_EVENT {YES, NO}

@DATA

01, HISTORICAL_BACKGROUND, NO, NO, NO, NO, **YES**, NO, NO, NO, **YES**, NO, NO, NO, NO, NO, NO, NO, NO, NO, NO, NO, **YES**, NO, NO, NO, NO, NO, NO, NO, NO, NO

More precisely, the Header describes the list of attributes. The format of @attribute is "@attribute [attribute-name] [values]". In our case, we have 31 features or attributes. The first one (@attribute PAIR) codifies the numeric id of each sentence pair of the corpus. The second attribute (@attribute RELATION) describes the CST relations of complementarity, which means that it has 3 possible nominal values: elaboration, historical_background, and follow-up. Then, there is an attribute for each of the 29 specific signals of the typology. Each of them was codified as a binary feature, admitting 2 possible values: "yes" for the presence and "no" for the absence of a signal.

The order the attributes are declared indicates the column position in the Data section, which describes the corpus examples (or instances) for training. If an attribute is the third one declared, then Weka expects that all those attribute values will be found in the third comma delimited column. As an illustration, consider the instance in the Data section of Figure 4, which corresponds to example (4). According to the order of the attributes in the Header, the 3 signals of complementarity that occur in this example (i.e., morphological=preposition, syntactic=superlative, and semantic=named entity) (see Figure 3), for instance, are respectively found in the 7th, 11th and 23rd comma delimited columns, which is indicated by the value "yes". The null occurrence of any referential or even pragmatic signal in example (4) is indicated by the value "no".

To perform the ML over the complementarity subcorpus from CSTNews, we applied the *10-fold cross-validation* technique. In the basic *10-fold cross-validation*, the corpus is randomly partitioned into *k* equal sized subsamples. Of the *k* subsamples, a single one is retained for test, and the remaining (*k – 1*) subsamples are used as training data. The process is repeated *k* times, with each of the k subsamples used once as the test data. The results are averaged over all the runs. We selected the *10-fold cross-validation* method because it gets more realistic estimates of the error rates for

classification, since our dataset is relatively small (i.e., 655 pairs of sentences) and unbalanced (i.e., 76 of *Historical background*, 260 of *Follow-up*, and 319 of *Elaboration*).

Additionally, we also performed the attribute selection process using Weka. Even with a relatively small set of attributes from the typology (i.e., 29 features, precisely), this selection is an ML technique that reveals the importance of the features, reducing processing time as well as increasing the performance of mining task.

Although there are different ML paradigms available in Weka, i.e., connectionist, mathematical (or probabilistic) and symbolic, we focused on symbolic algorithms, because they produce rules that can be easily interpreted and verified by human experts. Nonetheless, we have also tested other machine algorithms from other Artificial Intelligence paradigms, for comparison purposes only (Table 7).

To evaluate the results, we have used the following metrics: accuracy, precision (P), recall (R), and *f*-measure (f-m). Accuracy indicates an overall performance of the model or classifier; such metric determines how far the output can be from the optimal one). Precision is the percentage value indicating how many of the instances returned by the algorithm are correctly classified. Recall quantifies the number of positive class predictions made out of all positive examples in the dataset. F-Measure provides a single score that balances both the concerns of precision and recall in one number, being a unique indicator of the algorithm performance (SHALEV-SHWARTZ, BEN-DAVID, 2014; JURAFSKY, MARTIN, 2021).

Table 7 – Automatic validation of the signal typology.

Paradigms / **CST relations**	**Symbolic**						**Mathematical**			**Connectionist**		
	JRip (96.6%)			**J48 (95.8%)**			**NaïveBayes (96.7%)**			**MLP (96.1%)**		
	P	**R**	***f-m***	**P**	**R**	***f-m***	**P**	**R**	***f-m***	**P**	**R**	***f-m***
HB	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.99	1.00	1.00	1.00
Follow-up	0.96	0.92	0.94	0.98	0.93	0.95	0.98	0.94	0.96	0.95	0.94	0.95
Elaboration	0.94	0.97	0.95	0.94	0.96	0.97	0.95	0.98	0.96	0.95	0.96	0.96

According to Table 7, the algorithms from different paradigms present very similar general accuracy. The well-known connectionist Multi-Layer Perception (MLP) algorithm achieved 96.1% of general accuracy with the default Weka configurations. Among the several mathematical or probabilistic methods in Weka, we ran Naïve-Bayes, which presents the highest general accuracy among all the algorithms, achieving 96.7%. More specifically, we tried JRip and J48 algorithms from the symbolic paradigm. JRip and J48 generated sets of rules with 96.6% and 95.8% of accuracy, respectively. The decision tree produced by J48, however, has more rules than the JRip classifier (22 and 9 rules, respectively).

More than a good classification accuracy, we wanted to be able to make the characterization of the different CST relations (of complementarity) explicit. Thus, we explored the results achieved by the symbolic algorithms in more detail. We chose to use the JRip classifier based on the combination of two factors: (i) manageable rule set and (ii) highest accuracy among the symbolic algorithms. We believe that combination is a good scenario for our purposes in this work. In other words, we chose JRip because its classifier learned a small set of rules with the best general accuracy

We present in Table 8 the 9 rules of the JRip algorithm, which are followed by the number of instances (sentence pairs) correctly (C) classified and incorrectly (I) classified, and the precision of the rule. For example, the 1^{rst} rule of Table 7 predicated a total of 38 pairs, and all of them were correctly classified as *Historical background,* thus it has 100% precision. In the rules, one can say that aspectuality is a signal type (from the pragmatic class) that characterizes well the *Historical Background* relation, since the 1^{rst} and 2nd rules are based on durative and punctual events, respectively. This means that the historical context is commonly an event/fact that occurs frequently or a specific

event back in time. By the way, the attribute selection using the InfoGainAttributeEval[6] algorithm (at Weka) also indicated the relevance of the aspectuality feature.

Table 8 – JRip logic rules.

Rule	C	I	P (%)
1. If DurativeEvent=*yes* then *Historical background*	38	0	100
2. Else-if PunctualEvent=*yes* then *Historical background*	38	0	100
3. Else-if Tense=*yes* and Theme-Rheme=*no* and RelatedInfo=*no* then *Follow-up*	136	2	98.5
4. Else-if NamedEntity(Time)=*yes* then *Follow-up*	58	3	95
5. Else-if Posteriority=*yes* then *Follow-up*	39	0	100
6. Else-if Continuity=*yes* then *Follow-up*	11	0	100
7. Else-if RelatedWords=*yes* and Numeral=*yes* then *Follow-up*	2	0	100
8. Else-if Detailing=*yes* and ElocVerb=*yes* and NomAnaphor=*yes* then *Follow-up*	8	2	80
9. Else-if *Elaboration*	325	13	96.6

The *Follow-up* relation is characterized by signals from all classes (referential, morphological, syntactic, semantic and pragmatic), since all rules, from the 3rd to the 8th, are based on them. Rules four, five and six are based on individual signals, while rules three, seven and eight are based on combinations of specific signals. The 3rd rule, for example, combines 3 features from different signal types/classes to classify the higher number of *Follow-up* instances: (i) verbal tense (word class→morphological), (ii) theme-rheme (displacement→syntactic class) and (iii) related information (addition→pragmatic class). In rule five, we see that posteriority is the signal that characterizes individually the higher number of *Follow-up* instances (39 sentences pairs) with 100% precision.

[6] The algorithm evaluates the worth of a feature by measuring the information gain with respect to the class.

If none of the eight first rules are applied, the default class is *Elaboration*, which is given by the 9th rule. This could indicate that *Elaboration* is not characterized by particular signals present on this set of attributes, being a very generic CST relation.

5 Final remarks and future works

We have presented a validation task of the signaling taxonomy of complementarity proposed by Souza (2019). The purpose of this work was to determine to what extent complementarity carry textual signals that may help NPL applications identify the correspondent CST relations. NLP research so far has focused mainly on temporal complementarity (i.e., *Historical background* and *Follow-up* relations) (SOUZA; DI-FELIPPO, 2018), but complementary information not related to temporal attributes is very frequent in language. This can be seen in CSTNews, where *Elaboration* is the second most frequent relation (20.90%) in the corpus, corresponding to 48.10% of the total CST relations of complementarity. Thus, it is essential to explore automatic ways for identifying both.

In the process of annotating or describing the sentence pairs in the subcorpus of temporal and non-temporal complementarity, Souza and Di-Felippo (2018) and Souza (2019) have noticed that delimiting and classifying the signals are not easy tasks, given the subjectivity involved in the task. Additionally, the ML study reveals the relevance of the signals for distinguishing the different CST relations in question. In this respect, we can confidently say that signals can potentially support the automatic detection of the relations because the JRip´s classifier had 96.9% of accuracy with a small set of rules.

However, it is important to note that not all signals from the typology are machine-tractable attributes, mainly those from the pragmatic class. This means that there are no NLP tools for automatically annotating them in corpora. Thus, one future goal is to investigate only machine-treatable signals using the whole subcorpus of

complementarity (i.e., 655 pairs). This is the case of the morphological and syntactic signals, for example, which can be identified by taggers and parsers, respectively.

We emphasize the contributions of this work from two natural instances to this field of study: (i) *Descriptive Linguistics* and (ii) *NLP*. In (i), when we systematize a broad set of linguistic signals of complementarity that expand the linguistic knowledge that we had until then about the phenomenon; and in (ii) by providing subsidies (linguistic signals) for the automatic identification of complementarity, one of the most frequent linguistic phenomena in multi-document journalistic corpora, and whose identification can help in the task of automatic summarization.

Finally, our work can help to enrich CSTNews, which is the reference corpus for MDS. The delimitation and description tasks of signals (see example (4)) can be used to insert a new type or layer of annotation to the corpus. We believe that this type of linguistic annotation may be used in future research on multi-document analysis.

Acknowledgment

We thank the Coordination for the Improvement of Higher Education Personnel for the financial support and the Interinstitutional Center of Computational Linguistics for the research support.

References

ALEIXO, P.; PARDO, T.A.S. Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts. *In*: **Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web**. 2008. p. 298-303. DOI https://doi.org/10.1145/1809980.1810055

BELTRAME, W.; CURY, D.; MENEZES, C. S. Fique Sabendo: um Sistema de Disseminação Seletiva da Informação para Apoio à Aprendizagem. *In*: **Brazilian symposium on Computers in Education**. Rio de Janeiro – Brazil. 2012. 10p.

CARDOSO, P. C. F.; MAZIERO, E. G.; JORGE, M. L. C.; SENO, E. M. R.; DI-FELIPPO, A.; RINO, L. H. M.; NUNES, M. G. V.; PARDO, T. A. S. CSTNews: a discourse-annotated corpus for Single and Multi-Document Summarization of news texts in Brazilian Portuguese. *In*: **Proceedings of the 3rd RST Brazilian Meeting**. Cuiabá – Brazil. 2011. p. 88-105.

DAS, D.; TABOADA, M. RST signalling corpus: a corpus of signals of coherence relations. **Language Resources and Evaluation**, v. 52, n. 1, p. 149–184, 2018. DOI https://doi.org/10.1007/s10579-017-9383-x

INAM, S.; SHOAIB, M.; MAJEED, F.; SHAERJEEL, M. I. Ontology based query reformulation using rhetorical relations. **International Journal of Computer Sciences IJCS**, Vol 9, Issue 4. p. 261-268, 2012.

JURAFSKY, D; MARTIN, J. **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3ª Edition (Draft), 2021. Available at: https://web.stanford.edu/~jurafsky/slp3/. Access in: 08 Sept. 2021.

KUMAR, Y. J.; SALIM, N.; RAZA, B. Cross-document structural relationship identification using supervised machine learning. **Applied Soft Computing**, v. 12, n. 10, p. 3124-3131, 2012. DOI https://doi.org/10.1016/j.asoc.2012.06.017

MANI, I. **Automatic summarization**. Vol. 3. John Benjamins Publishing. 2001. DOI https://doi.org/10.1075/nlp.3

MANN, W. C.; THOMPSON, S. A. **Rhetorical structure theory**: A theory of text organization. University of Southern California, Information Sciences Institute, 1987. DOI https://doi.org/10.1515/text.1.1988.8.3.243

MAZIERO, E. G.; JORGE, M. L. C.; PARDO, T. A. S. Identifying multi-document relations. *In*: **Proceedings of the International Workshop on Natural Language Processing and Cognitive Science**. Funchal, Madeira/Funchal. 2010. p. 60-69.

MAZIERO, E.; PARDO, T. A. CSTParser–a multi-document discourse parser. *In*: **Proceedings of the PROPOR**. Coimbra – Portugal. 2012. p. 1-3.

MAZIERO, E. G. **Identificação automática de relações multidocumento**. Master's dissertation (Masters in Computer Science and Computational Mathematics) - Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, 2012.

MAZIERO, E. G.; JORGE, M. L. R. C.; PARDO, T. A. S. Revisiting Cross-document Structure Theory for multi-document discourse parsing. **Information Processing & Management**, v. 50, n. 2. p. 297-314, 2014. DOI https://doi.org/10.1016/j.ipm.2013.12.003

MITCHELL, T. M. Does machine learning really work? **AI magazine**, v. 18, n. 3, p. 11. 1997.

MURAKAMI, K.; NICHOLS, E.; MIZUNO, J.; WATANABE, Y.; GOTO, H.; OHKI, M. Automatic classification of semantic relations between facts and opinions. *In*: **Proceedings of 2nd workshop on NLP challenges in the information explosion Era NLPIX**. Beijing – China. 2010. p. 21–30.

NENKOVA, A.; MCKEOWN, K. Automatic summarization. **Foundations and Trends in Information Retrieval**, 5(2-3), p. 103–233, 2011. DOI https://doi.org/10.1561/1500000015

RADEV, D. R. A. Common theory of information fusion from multiple text sources step one: cross-document structure. *In*: **Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue**. Volume 10. 2000. p. 74-83. DOI https://doi.org/10.3115/1117736.1117745

SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding Machine Learning**: From Theory to Algorithms. New York: Cambridge University Press, 2014. DOI https://doi.org/10.1017/CBO9781107298019

SOUZA, J. W. C. **Descrição linguística da complementaridade para a sumarização automática multidocumento**. *Dissertação* (Mestrado em Linguística) – Universidade Federal de São Carlos. 2015. p. 102.

SOUZA, J. W. C.; DI-FELIPPO, A. Caracterização da complementaridade temporal: subsídios para sumarização automática multidocumento. **Alfa: Revista de Linguística** (São José do Rio Preto), v. 62, p. 125-150, 2018. DOI https://doi.org/10.1590/1981-5794-1804-6

SOUZA, J. W. C. **Aprofundamento da caracterização linguístico-computacional da complementaridade em um corpus jornalístico multidocumento**. 2019. *Tese* (Doutorado em Linguística) – Universidade Federal de São Carlos, São Carlos, p. 117. 2019.

SOUZA, J. W. C. O papel do corpus de estudo no aprimoramento descritivo da complementaridade informacional multidocumento. **Revista de Estudos da Linguagem**, v. 29, n. 2, 2021. DOI https://doi.org/10.17851/2237-2083.29.2.1059-1087

TABOADA, M.; DAS, D. Annotation upon annotation: adding signalling information to a corpus of discourse relations. **Dialogue and Discourse**. v. 4, n. 2, p. 249-281, 2013. DOI https://doi.org/10.5087/dad.2013.211

WITTEN, I. H.; FRANK, E. **Data Mining**: Practical machine learning tools and techniques. 2nd edition. Morgan Kaufmann, San Francisco. 2005.

ZAHRI, N. A. H. B.; FUKUMOTO, F. Multi-document Summarization using link analysis based on rhetorical relations between sentences. *In*: **CICling Lectures Notes in Computer Science**. 2011. p. 328-338. DOI https://doi.org/10.1007/978-3-642-19437-5_27

ZHANG, Z.; BLAIR-GOLDENSOHN, S.; RADEV, D. R. Towards CST-enhanced summarization. *In*: **Proceedings of the 18th National Conference on Artificial Intelligence** (AAAI-2002), Edmonton – Canada. 2002. p. 439-446.

ZHANG, Z.; OTTERBACHER, J.; RADEV, D. R. Learning cross-document structural relationships using boosting. *In*: **Proceedings of 12th ICIKM**. New Orleans, USA. 2003. p. 124–130. DOI https://doi.org/10.1145/956863.956887

Article received in: 10.26.2021 Article approved in: 04.28.2022