Subtopic annotation and automatic segmentation for news texts in Brazilian Portuguese

Paula C.F. Cardoso[1], Thiago A.S. Pardo[2] and Maite Taboada[3]

Abstract

Subtopic segmentation aims to break documents into subtopical text passages, which develop a main topic in a text. Being capable of automatically detecting subtopics is very useful for several Natural Language Processing applications. For instance, in automatic summarisation, having the subtopics at hand enables the production of summaries with good subtopic coverage. Given the usefulness of subtopic segmentation, it is common to assemble a reference-annotated corpus that supports the study of the envisioned phenomena and the development and evaluation of systems. In this paper, we describe the subtopic annotation process in a corpus of news texts written in Brazilian Portuguese, following a systematic annotation process and answering the main research questions when performing corpus annotation. Based on this corpus, we propose novel methods for subtopic segmentation following patterns of discourse organisation, specifically using Rhetorical Structure Theory. We show that discourse structures mirror the subtopic changes in news texts. An important outcome of this work is the freely available annotated corpus, which, to the best of our knowledge, is the only one for Portuguese. We demonstrate that some discourse knowledge may significantly help to find boundaries automatically in a text. In particular, the relation type and the level of the tree structure are important features.

Keywords: corpus annotation, newspaper discourse, subtopics, text segmentation.

[1] Department of Computer Science, Federal University of Lavras, 37200-000, Lavras, Brazil.
[2] Department of Computer Science, University of São Paulo, 13566-590, São Carlos, Brazil.
[3] Department of Linguistics, Simon Fraser University, 8888 University Dr., Burnaby, B.C. V5A 1S6, Canada.
Correspondence to: Paula C.F. Cardoso, *e-mail*: paulastm@gmail.com

Corpora 2017 Vol. 12 (1): 23–54
DOI: 10.3366/cor.2017.0108

www.euppublishing.com/cor

1. Introduction

Subtopic segmentation aims to find the boundaries among subtopic blocks in a text (Hearst, 1997). This task is useful for many important applications in Natural Language Processing (NLP), such as automatic summarisation, question answering, and information retrieval and extraction. For instance, Wan (2008) states that, given some subtopic segmentation, automatic summarisation may produce summaries that select different aspects from the collection of texts, thus resulting in better summaries. Oh *et al.* (2007) suggest that a question answering system, which aims to answer a question/query submitted by the user, may link this query to the subtopics in a text in order to increase the accuracy of the identification of the answer. Prince (2007) explains that information retrieval with the identification of subtopics in the retrieved texts may provide the user with text fragments that are semantically and topically related to a given query. This makes it easier for the user to find quickly the information that is of interest.

To illustrate, Figure 1 (translated from the original language, Portuguese) shows a short text with sentences identified by numbers between square brackets, and a possible segmentation. We also show the identification of each subtopic in angle brackets after the corresponding text passages. The main topic is the health of Maradona, a famous Argentinian soccer player. The first block discusses Maradona's relapse due to hepatitis, the second block describes his current state of health, and the last one reports that he received messages of support from fans.

As in other studies (Bollegala *et al.*, 2006; Du *et al.*, 2013; Hearst, 1997; Hennig, 2009; Hovy, 2009; and Riedl and Biemann, 2012), we assume that a text or a set of texts develop a main topic, exposing several subtopics (pieces of text that cover different aspects of the main topic) as well. For example, a set of news texts related to an earthquake typically contains information about the magnitude of the earthquake, its location, casualties and rescue efforts (Bollegala *et al.*, 2006). Therefore, the task of subtopic segmentation aims to divide a text into topically coherent segments. Several methods have been tested for subtopic segmentation (Chang and Lee, 2003; Choi, 2000; Du *et al.*, 2013; Hearst, 1997; Hovy and Lin, 1998; Passonneau and Litman, 1997; and Riedl and Biemann, 2012). However, there are no studies on how discourse structure mirrors subtopic boundaries in texts and how they may contribute to such a task, although such possible correlation has been suggested (e.g., Hovy and Lin, 1998).

For segmenting texts by subtopic, it is important to prepare a reference segmentation that supports not only the study and understanding of the phenomenon, but also the development and evaluation of systems for automatic subtopic segmentation. As the construction of corpora is a time consuming and expensive task, Hovy and Lavid (2010) recommend that it is necessary to be concerned with the reliability, validity and consistency of the corpus annotation process, in order to produce a scientifically sound

[S1] Maradona had again health problems over the weekend.
[S2] Hospitalized in Buenos Aires, he had a relapse and felt pain again due to acute hepatitis, according to his personal doctor, Alfredo Cahe.
<subtopic: Maradona's relapse>

[S3] "Nos his state of health is stable. Despite this improvement, he is still hospitalized", said the doctor, who has discarded the possibility that the ex-player has pancreatitis (inflammation of the pancreas, an organ located behind the stomach and that influences the digestion).
[S4] Cahe emphasized that Maradona still has problems.
[S5] His liver values are not balanced and he is not well. But it is nothing serious", he said in an interview for the la Nación newspaper.
<subtopic: current state of health>

[S6] On Sunday, Maradona watched the 1-1 draw in the classic Boca Juniors and River Plate on television.
[S7] Boca Junior's fans, who turned out in large number to the stadium La Bombonera, led many banners and flags with messages of support for the Argentinian idol.
[S8] His daughter, Dalma, was in the stadium to watch the game.
<subtopic: messages of support>

[S1] Maradona voltou a ter problemas de saúde no fim de semana.
[S2] Internado em um hospital em Buenos Aires, ele teve uma recaída e voltou a sentir dores devido a hepatite aguda que o atinge, segundo seu médico pessoal, Alfredo Cahe.
<subtópico: recaída>

[S3] "Agora está estável. Mesmo com esta melhora, ele continuará internado", disse o médico, que descartou a possibilidade do ex-jogador ter uma pancreatite (inflamação do pâncreas, órgão situado atrás do estômago e que influencia na digestão).
[S4] Cahe reforçou que Maradona ainda tem problemas.
[S5] "Os valores hepáticos dele na avaliação não estão equilibrados e ele não está bem. Mas não é nada grave", afirma, em entrevista ao diário La Nación.
<subtópico: estado atual>

[S6] No domingo, Maradona assistiu ao empate por 1 a 1 no clássico Boca Juniors e River Plate pela televisão.
[S7] Os torcedores do Boca, que compareceram em grande número ao Estádio La Bombonera, levaram muitas faixas e bandeiras com mensagens de apoio ao ídolo argentino.
[S8] Sua filha, Dalma, foi ao estádio assistir ao jogo.
<subtópico: mensagens de apoio>

Figure 1: Example of a text segmented in subtopics.

resource. Because of this, many researchers (for instance, Aluísio *et al.*, 2014; da Cunha *et al.*, 2011; and Iruskieta *et al.*, 2013) have followed the steps proposed by Hovy and Lavid (2010): (*1*) choosing the phenomenon to annotate and the underlying theory, (*2*) selecting the appropriate corpus, (*3*) selecting and training the annotators, (*4*) specifying the annotation procedure, (*5*) designing the annotation interface, (*6*) choosing and applying the evaluation measures, and (*7*) delivering and maintaining the product.

In this paper, we report the subtopic annotation of news texts written in Brazilian Portuguese, following these seven steps that represent the major research questions in corpus annotation. The corpus, called CSTNews (Cardoso *et al.*, 2011), was originally designed for multi-document processing and contains fifty clusters, with each cluster having two or three texts on the same topic.

Using this corpus, we have then developed and evaluated methods for subtopic segmentation of news texts. In particular, we were driven by the belief that discourse organisation mirrors subtopic changes in the texts. Therefore, our methods rely mainly on discourse features, and we explore the potential of Rhetorical Structure Theory (RST; Mann and Thompson, 1987) for this purpose. We compare our results to some well-known algorithms in the area and show that our proposal out-performs them, evidencing the usefulness of discourse for the task of subtopic segmentation and, in more general terms, subtopic modelling.

To the best of our knowledge, the corpus annotated with subtopic segments is the first one to be available for Portuguese. At the same time, the correspondence between discourse and subtopic structures is systematically shown here for the first time.

The remainder of this paper is organised as follows. In the next section, we give an overview of existing corpora annotated with subtopics and some well-known algorithms for subtopic segmentation, as well as a brief introduction to RST. We then report our corpus annotation, following the steps (methodology) proposed by Hovy and Lavid (2010). Next, we describe our automatic strategies to find subtopic boundaries, which are followed by an evaluation and discussion of the results. Finally, we present conclusions and present some considerations for future work.

2. Background

2.1 Existing corpora

There are several initiatives to create corpora that are linguistically annotated with varied phenomena from diverse perspectives, both for written and spoken/transcribed data. We briefly overview some of these works for subtopic segmentation, below.

Hearst (1997) used a corpus of twelve magazine (expository) articles that had their subtopics annotated by seven technical researchers. This kind of text consists of long sequences of paragraphs with very little structural demarcation. In order to produce a reference annotation (reference segmentation), the author considered that a subtopic boundary occurred if at least three out of the seven judges placed a boundary mark there. The annotators were simply asked to mark the paragraph boundaries at which the subtopic changed; they were not given more explicit instructions about the granularity of the segmentation.

Annotator agreement is an important measure in corpus annotation to show how well the annotators understand the phenomenon at hand, how much they agree on the data and, therefore, how reliable the corpus annotation is, allowing the development and evaluation of systems and theories. Hearst (1997) and others used the traditional kappa (k) measure (Carletta, 1996) to compute agreement among annotators. The kappa measure produces results in the range of 0 to 1, with 1 indicating perfect agreement. Carletta (1996) states that $k > 0.8$ shows good replicability/reliability, and $0.67 < k < 0.8$ allows tentative conclusions to be drawn; it is also known that such values highly depend on the subjectivity of the task at hand and, for more subjective annotations, lower values may well be accepted (as is usually the case for subtopic segmentation). In Hearst's annotation, the agreement among annotators was 0.64.

Kazantseva and Szpakowicz (2012), in turn, chose a fiction book (with twenty chapters) to be segmented by at least four undergraduate students. The annotators were divided into five groups and each group read and annotated four distinct chapters. Each annotator worked individually. For each subtopic boundary, the annotator provided a brief one-sentence description, effectively creating a chapter outline. The authors also chose to use paragraphs as the basic unit of annotation. The analysis of the resulting corpus revealed that the overall inter-annotator agreement was low ($k = 0.29$) and was not uniform throughout the chapters. The authors decided not to provide a reference segmentation.

Passonneau and Litman (1997) used a corpus consisting of twenty transcribed narratives about a movie to be annotated by seven untrained annotators. The authors asked judges to mark boundaries using their notion of communicative intention as the segmentation criterion. Judges were also instructed to identify briefly the speaker's intention associated with each segment. For reference segmentation, it was necessary that at least four out of the seven judges placed a boundary mark in the corresponding point of the text. The authors report that the agreement was 60 percent on boundaries.

Galley *et al.* (2003) worked on a sample of twenty-five meetings transcribed from the ICSI Meeting corpus (Janin *et al.*, 2003). The transcripts detailed the speaker, start time, end time and content for each participant. The authors had at least three human judges mark each speaker change (which is a potential boundary) as either a boundary or non-boundary. The final segmentation was based on the opinion of the majority.

Gruenstein *et al.* (2007) used forty transcribed meetings from the same ICSI Meeting corpus and sixty additional ones from the ISL Meeting Corpus (Burger *et al.*, 2002). The authors asked two annotators to segment the texts at two levels: major and minor, corresponding to the more and less important subtopic shifts. Annotators also gave brief descriptive names to subtopics. Kappa values were 0.47 for major subtopics, and 0.46 for minor subtopics. The authors noticed many cases where subtopic boundaries were annotated as a major shift by one annotator and as a minor shift by another, leading to low agreement. Again, reference segmentation was not provided.

Other researchers automatically produced reference segmentation data. For instance, Choi (2000) produced an artificial test corpus of 700 documents from the Brown corpus (Francis and Kučera, 1979). For document generation, the procedure consists of extracting, for instance, ten segments of three to eleven sentences each, taken from different documents and combining them to form one document.

Considering these related works, some issues may be observed. For instance, the agreement among annotators is generally low. The researchers report that the main difference in the annotations is the granularity: some annotators mark only the most prominent boundaries (coarse granularity), while others find finer changes (smaller segments). On the other hand, even though humans vary widely, some agreement may be found and is sufficient to demonstrate that the annotators are consistent in relation to major subtopics. The minimum number of annotators in agreement was three, and a reference segmentation is based on the majority opinion. In general, most of the previous works asked the annotators to put a mark where there was a boundary and to include a brief description. None of these papers report how researchers make use of the subtopic descriptions provided by the annotators. It is possible that these descriptions served to show the researchers whether participants had understood the phenomenon that was being investigated.

We can also notice how versatile these annotation procedures are. This is expected, since corpora are created for several different (linguistic and computational) purposes. However, corpus annotation practices have evolved with time and some basic steps are expected to be followed in the research. As already mentioned, Hovy and Lavid (2010) split the corpus annotation into seven steps and argue that it is necessary to follow them in order to have reliable corpora and, therefore, trustworthy applications. We follow these steps in our work, and we also base our main annotation decisions on some of the above works on corpus annotation, as will be shown later in this paper.

2.2 Methods for automatic text segmentation

Several approaches have been developed for subtopic segmentation. They usually measure similarity across sentences and place subtopic boundaries where the similarity between adjacent sentences, windows of words or paragraphs is low. One well-known approach, which is widely used for subtopic segmentation, is TextTiling (Hearst, 1997), which is based on lexical cohesion. For this strategy, it is assumed that a set of lexical items is used during the development of a subtopic in a text and, when that subtopic changes, a significant proportion of vocabulary also changes. For identifying major subtopic shifts, text passages of pre-defined size k (blocks) are compared for overall similarity. The more words these blocks have in common, the higher the probability that they address the same subtopic. Hearst (1993, 1994, 1997) wrote that there were several ways that the TextTiling algorithm could be modified. One of the adjustable parameters is the size of the block used for comparison. The k value slightly varies from text to text; as a heuristic, it is assigned to k the average paragraph length (in sentences). According to the author, TextTiling was evaluated with twelve magazine articles and achieved 71 percent precision and 59 percent recall on boundaries.

Choi (2000) developed the algorithm called C99, also based on lexical cohesion. Starting from pre-processed sentences, C99 initially calculates the similarity between each pair of sentences and produces a similarity matrix. From the matrix, it produces a rank-similarity: the more similar the sentences are with their neighbours, the higher the score will be. The lower ranks in the classification matrix indicate subtopic boundaries.

Passonneau and Litman (1997), in turn, have combined multiple linguistic features for subtopic segmentation of spoken text, such as pause, cue words and referential noun phrases. The evaluation showed that the noun phrase feature performs better than the others, with 50 percent recall and 31 percent precision. The authors relate this to the fact that noun phrases encompass more knowledge than pause and cue-word features. These two features (pause with 92 percent recall and 18 percent precision; and cue words with 72 percent recall and 15 percent precision) assigned many boundaries, but they were not in accordance with those specified by most of the human annotators. The authors argue that text segmentation involves much more than using shallow linguistic knowledge and other possibilities ought to be investigated.

Riedl and Biemann (2012), based on TextTiling, proposed the TopicTiling algorithm that segments documents using the Latent Dirichlet Allocation (LDA) topic model (Blei *et al.*, 2003). The documents that are to be segmented have first to be annotated with topic IDs, obtained by the LDA inference method. The topic model must be trained on documents similar in content to the test documents. The IDs are used to calculate the cosine similarity between two adjacent sentence blocks, represented as two vectors, containing the frequency of each topic ID. Values close to 0 indicate marginal relatedness between two adjacent blocks, whereas values close to 1 denote connectivity. For evaluating, the authors applied WindowDiff measure (Pevzner and Hearst, 2002): the results have shown that TopicTiling improves the state of the art.

Du *et al.* (2013) presented a hierarchical Bayesian model for unsupervised topic segmentation. The model takes advantage of the high modelling accuracy of structured topic models to produce a topic segmentation based on the distribution of latent topics. The model consists of two steps: modelling topic boundary and modelling topic structure. The authors evaluated the algorithm on three different kinds of corpora: a set of synthetic documents, two meeting transcripts and two sets of text books. The model shows prominent segmentation performance on either written or spoken texts using WindowDiff.

Hovy and Lin (1998) have used various complementary techniques, including those based on text structure, cue words and high-frequency indicative phrases for subtopic identification in a summarisation system.

Although the authors do not mention an evaluation of these features, they argue that discourse structure might help subtopic identification.

In this section, different segmentation strategies were presented. As we can see, none of them follows the discourse structure as suggested by Hovy and Lin (1998). We also believe it is possible to find the structure of subtopics by exploring the discourse organisation of a text. For this, Hovy and Lin (1998) suggested using RST (Mann and Thompson, 1987, 1988), which is the strategy that we follow here. Therefore, RST is briefly introduced, below.

3. Rhetorical Structure Theory

Rhetorical Structure Theory (RST) represents relations among propositions or discourse segments/spans in a text and differentiates between nuclear and satellite information (Mann and Thompson, 1987, 1988). Nuclei are considered to be the most important parts of a text, whereas satellites contribute to the nuclei and are secondary. The distinction between nuclei and satellites comes from the observation that the nucleus is more essential to the writer's purpose than the satellite. The satellite is often incomprehensible without the nucleus, whereas a text where the satellites have been deleted can be understood to a certain extent (Taboada and Mann, 2006). In order to present the differences among relations, they are organised in two groups: subject matter and presentational relations. In the former, the text producer intends that the reader recognises the relation itself and the conveyed information (e.g., CONTRAST or ELABORATION), while in the latter the intended effect is to increase some inclination on the part of the reader (e.g., ANTITHESIS or JUSTIFY) (Taboada and Mann, 2006). Relations with one nucleus and one satellite are referred to as mononuclear relations. Relations that only have nuclei (where all the propositions are equally important) are said to be multinuclear relations.

The relations are structured in a tree-like form (where larger units – consisting of more than one proposition – are also related in the higher levels of the tree). As an example of a rhetorical analysis of a text, consider the already segmented text and its rhetorical structure shown in Figure 2. The symbols *N* and *S* indicate the nucleus and the satellite of each rhetorical relation. In this structure, Propositions 3 and 4 are comparable items linked by a LIST relation; this subtree is the entire nucleus of the RESULT relation and they could be causes for the situation presented in Proposition 2; this whole subtree is the nucleus of the ATTRIBUTION relation. One may notice that the result of organising a text based upon RST is a hierarchical structure, and leaves are text spans which correspond to the propositions, or, as named by Marcu (2000a), the Elementary Discourse Units (EDUs) in a text. A more detailed explanation of RST may be found in Mann and Thompson (1987, 1988).

Next, we present our systematic corpus annotation, following the seven steps proposed by Hovy and Lavid (2010).

[1]According to police information, [2]Abadia was arrested in March 1996 [3]for sending to the United States 30 tons of cocaine [4]and sending drugs to the country in association with a cartel of Tijuana, Mexico.

[1] De acordo com a polícia, [2]Abadia foi preso em março de 1996 [3]por enviar aos Estados Unidos 30 toneladas de cocaína e [4]mandar drogas ao país em associação com um cartel de Tijuana, no México.

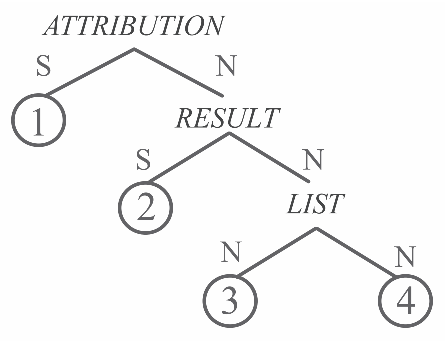

Figure 2: Example of Rhetorical Structure.

4. Subtopic annotation in the CSTNews corpus

4.1 The decision on which linguistic phenomenon to annotate

The phenomenon we focus on is subtopic segmentation in news texts. According to Koch (2009), a text can be regarded as coherent if it displays continuity (i.e., subtopic progression must take place in such a way that there are no overly long breaks or interruptions of the subtopic in progress). As described in the first section, we assume that a text develops a main topic, exposing several subtopics as well, which make sense together (showing continuity, as a consequence) and contribute to the overall topic.

The structure of subtopics is usually marked in technical reports and scientific texts by headers and sections that divide the text into coherent segments. News texts, letters and blogs do not usually have explicit marking of subtopics; however, the structure of subtopics exists as an organising principle of the text.

It is usual to find subtopics that are repeated later on in the same text, and should, therefore, be connected and identified with the same label. It indicates that a subtopic that was already described may return after a certain time, which means that sentences that belong in the same subtopic are not always adjacent. In addition, the granularity of a subtopic is not defined; and it may contain one or more sentences or paragraphs. Some researchers use paragraphs as the basic information unit (e.g., Kazantseva and Szpakowicz, 2012), while others employ sentences (e.g., Chang and Lee, 2003; and Riedl and Biemann, 2012).

We also do not distinguish between the notions of subtopic shift and subtopic drift, as proposed by Carlson and Marcu (2001). According to the authors, a subtopic shift is a sharp change in focus, while a drift is a smooth change from the information presented in the first segment to the information presented in the second. Let us consider the text of Figure 1 as an example. Between sentences S3 and S4, we could say there is a subtopic-drift relation because the two sentences describe Maradona's state of health at this moment. On the other hand, between sentences S3–S5 and S6, we say there is a subtopic-shift because S6 talks about a game. In our annotation, they were both simply classified as subtopic changes. We believe that distinguishing between subtopic shift and drift is a very subjective task and may possibly result in low agreement levels, as already observed in similar works. For example, Gruenstein *et al.* (2007) achieved low agreement when asked the annotators to segment texts at major and minor levels. Looking for signals in corpora that could help to identify relations in applications such as discourse parsing, Taboada and Das (2013) did not find any signals for topic-shift nor topic-drift, though they have found four instances of topic-shift. It is also interesting to notice that all the other works in the literature also opted not to tackle such issues.

4.2 Selecting the corpus

We used the CSTNews[4] corpus which is comprised of fifty clusters of news texts written in Brazilian Portuguese, manually collected from several important Brazilian news agencies, such as Folha de São Paulo, O Estadão, O Globo, Jornal do Brasil, and Gazeta do Povo. The texts belong to the Politics, Sports, World, Daily News, Money and Science sections. The corpus contains 140 texts altogether, amounting to 2,088 sentences and 47,240 words. On average, the corpus contains, for each cluster, two or three texts, 41.76 sentences and 944.8 words.

The choice for news texts was motivated by our interest in pursuing general-purpose subtopic segmentation. News texts usually make use of everyday language and are simple, and widely accessible to ordinary people (Lage, 2002). As mentioned earlier, such texts usually do not present structural demarcation, which means that automatic subtopic segmentation is necessary for other NLP tasks to be carried out in the text, such as automatic summarisation. At the same time, however, the lack of structural demarcation makes segmentation a challenging task.

Each cluster contains texts on the same topic, as the corpus was built to foster research in multi-document processing, mainly in multi-document summarisation. As the selection of texts to form the corpus was driven by

[4] See: http://www.icmc.usp.br/pessoas/taspardo/sucinto/cstnews.html.

	Precision	Recall
Simple Textual Segments	0.91	0.91
Complex Textual Segments	0.78	0.78
Nuclearity	0.78	0.78
Relations	0.66	0.66

Table 1: Precision and Recall average values of RST annotation.

which topics were current and more prominent at the time (because more prominent topics are likely to be covered by different news sources), the distribution of texts is not uniform among sections and agencies. For instance, some sections, such as World and Daily News, have far more texts than Science and Money sections. We consider that such differences are not relevant for the proposed task. Instead, we favour news language, regardless of the section that each text came from.

The corpus is also annotated with RST. As we will explore RST to find some strategies for automatic subtopic segmentation, we present a brief description of the RST annotation for the CSTNews corpus, which has been performed in Cardoso *et al.* (2011). RST annotation was performed by eight annotators, with knowledge of RST and some experience of annotation. Agreement between annotators was automatically computed using RSTeval (Maziero and Pardo, 2009), which is based on the work of Marcu (2000b). In this tool, given a set of RST trees, a tree must be selected as the ideal one and the others are compared to it based on four criteria: (*1*) simple textual segments; (*2*) complex textual segments (i.e., two or more segments related by some RST relation(s)); (*3*) nuclearity of every text segment; and (*4*) RST relation that holds between the segments. The well-known precision and recall metrics were computed for each RST tree to capture the degree of similarity among trees for the same text.

In order to illustrate the annotation agreement process, let us consider a cluster composed of four RST trees. First, a tree is selected as the ideal one and the other three are compared to it, considering the four criteria mentioned above. This process is repeated four times, so that each time a different tree is selected as optimal. Then, the average agreement values are calculated for each criterion. Table 1 shows these values for the corpus. (Note that precision and recall values are the same due to the way the comparison of trees was performed.) According to the results, the best values of agreement were achieved in the segmentation process (simple textual segments), computed before the annotators discuss the analysis. This is mainly due to the segmentation rules that make this task less subjective than the other tasks. As expected, the worst agreement values were obtained for the relations the annotators indicated. For more details, see Cardoso *et al.* (2011).

For comparative purposes, using his original (similar) evaluation strategy, Marcu (2000b) reports numbers for a group of five texts annotated by two humans. Marcu obtained the following results: 0.88 precision and recall for simple textual segments; 0.90 precision and recall for complex textual units; 0.79 precision and 0.88 recall for nuclearity; and 0.83 precision and recall for relations. Da Cunha *et al*. (2011) also used the same method for evaluating the agreement in the annotation of the RST Spanish Treebank. Applied to eighty-four texts annotated by ten humans, the results were: 0.87 precision and 0.91 recall for simple textual segments; 0.86 and 0.87 for complex textual segments; 0.82 precision and 0.85 recall for nuclearity; 0.77 precision and 0.78 recall for relations. Although the texts, languages, and the amount of data used by these other authors and in this work are very different, such comparison gives an idea of the human ability to agree on the RST annotation process. In general, we consider that the agreement results in this work were satisfactory, given the subjectivity of the task and the fact that we are not far from the results obtained in other research.

4.3 Selecting and training the annotators

Our subtopic annotation was performed by fourteen annotators, all computational linguists with experience in corpus annotation, including undergraduate and graduate students, as well as professors.

The annotators went through training sessions during three days in one-hour daily meetings. During this step, the annotators were introduced to the task, its general rules and its relevance for NLP (for motivational purposes), segmented some news texts (which were not from the CSTNews corpus, in order to avoid bias in the actual annotation), and discussed their annotations, concepts and definitions related to the task, as well as compared their annotations. These training sections were conducted by two experienced annotators who had performed subtopic annotation on previous occasions.

We believe that all the annotators had some intuitive notion about the task. The three days of training proved to be enough for the annotators to acquire maturity in the process and to achieve satisfactory agreement. This agreement was empirically checked during discussions. As in other studies (Galley *et al*., 2003; Gruenstein *et al*., 2007; Hearst, 1997; Kazantseva and Szpakowick, 2012; and Passonneau and Litman, 1997), agreement among judges was not perfect, since there is a high degree of subjectivity in the task, given that it involves reading and interpreting texts. Overall agreement can, nevertheless, be clearly found.

4.4 Specifying the annotation procedure

The annotation phase took seven days in a daily one-hour meeting. Restricting the annotation meetings to one hour is a good strategy to avoid errors and inconsistencies in the annotation due to annotator fatigue. Performing the annotation in sequential days guarantees continuous enrolment and attendance of the annotators.

For every annotation day, the annotators were organised in two groups, with at least five annotators. The groups were randomly formed in each annotation session in order to avoid bias in the process. The annotators were instructed to read each text and to split it into subtopic blocks. The annotation was done individually and the participants were not allowed to discuss the task with one another. This process is similar to the one described by Hearst (1997).

The segmentation granularity was not defined. Paragraph boundaries should not be taken into account, since a subtopic may be described inside a paragraph with other subtopics or even in more than one paragraph. For each subtopic, each annotator was asked to identify it with a brief description using keywords. The description was inserted in a concise and representative tag such as ' < t label = "keywords" > '. It was not necessary to identify the main topic, since it is organised in subtopics. The boundaries identified by the majority of the annotators were assumed to be actual (reference) boundaries.

4.5 Annotation interface

Hovy and Lavid (2010) suggest that a computer interface designed for annotation must favour speed, at the same time avoiding bias in the annotation. Leech (2005), in turn, states that, although we may annotate texts using a general-purpose text editor, such a method may be slow and prone to errors, because the task has to be performed by hand.

Since there is not a large set of tags for our annotation (only one, in fact), we did not develop a specific interface. The annotators used their preferred text editor and added a tag whenever they found a subtopic boundary. Thus, the annotators had the ability to go back and look over the parts that they had already examined, and change markings if desired, because they were manipulating a tool they already knew.

On the other hand, allowing the annotators to use the editor they were more familiar with meant that we had to check the encoding (since different editors and operating systems may use varied encoding standards, such as Unicode and UTF-8), to make it uniform. Notepad was one example of a text editor that was used.

4.6 Choosing and applying evaluation measures

The underlying premise of an annotation is that if people cannot agree enough, then either the theory is wrong (or badly stated or instantiated), or the process itself is flawed (Hovy and Lavid, 2010). At first, it may seem

Day	Groups	Annotators	Texts per group	K
1	A B	6 7	10	0.656 0.566
2	A B	5 5	10	0.458 0.447
3	A B	7 5	10	0.515 0.638
4	A B	5 7	10	0.544 0.562
5	A B	5 5	10	0.643 0.528
6	A B	5 5	12 13	0.570 0.549
7	A	5	15	0.611
			Average	0.560

Table 2: Agreement values for subtopic segmentation.

intuitive to determine possible subtopic boundaries, but the task is subjective and levels of agreement among humans tend to be low (Hearst, 1997; Gruenstein *et al.*, 2003; Kazantseva and Spakowicz, 2012; and Passonneau and Litman, 1997). Agreement on subtopic annotation also varies depending on the text genre/type that is segmented. For example, technical reports have headings and sub-headings, while other genres, such as news texts, have little demarcation.

The quality of annotation may refer to the agreement and consistency with which it is applied. As adopted by Hearst (1997), we used the kappa measure (Carletta, 1996). From left to right, Table 2 shows the days of annotation, the groups of annotators (represented by the letters A and B), the number of annotators in each group, the number of texts that were annotated in each group per day, and the agreement values that we obtained.

All fourteen annotators were not able to attend every single annotation day; for example, in the first day thirteen out of fourteen annotators participated in the annotation. We may see that the first day produced the best agreement among annotators, with a 0.656 agreement value for Group A and 0.566 for Group B. On the other hand, the lowest agreement was in the second day, with 0.458 for Group A and 0.447 for Group B. It is difficult to explain the fluctuations according to the day, but it may be the case that, on the first day, the annotators benefitted from the recent training, and that, on the second day, their confidence faltered as they encountered new cases. Agreement measures after the second day seem to improve, with some fluctuation. The average agreement was 0.560. Although the value is considered to be satisfactory, it is a bit lower than the one obtained by Hearst (1997) in her seminal work, which was 0.647. This may be explained by

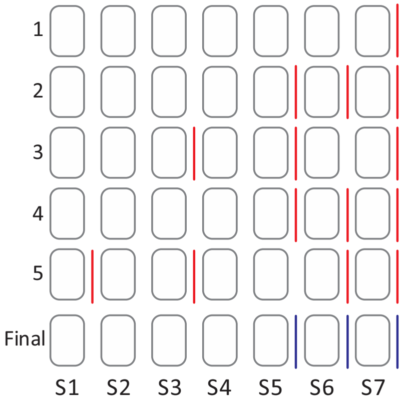

Figure 3: Example of different segmentations.

the fact that her work used fewer texts (only twelve), which were expository texts, where subtopic boundaries are usually more clearly indicated. Given the relatively low disagreement rate among annotators, we will argue for the reliability of the annotation procedure used in this paper.

From these annotated texts, a reference segmentation was established. We computed the opinion of the majority of the annotators (half plus one) in the boundaries. This strategy results in more reliable and coarse-grained subtopic boundaries. We observed only two texts for which the boundaries were in the same place for all annotators (not including the end of the text, since this happens for all of them). Other texts have higher or lower degrees of variation in segmentation. The variation in segmentation is related to factors such as the interpretation of the text and prior knowledge about the subject.

As an example of variation in segmentation, Figure 3 shows different segmentations for the text in Figure 4. The rows numbered from one to five represent the segmentation made by each of the five annotators. Each box represents a sentence and the segmentation is indicated by vertical lines. The last line, labelled ‘Final’, represents the reference segmentation, obtained from the majority of the annotators. One may see, for instance, that the first annotator did not place any subtopic boundary. The last boundary, after the last sentence, is expected, since the text ends. The second annotator placed boundaries after the fifth and sixth sentences, besides the one at the end of the text. The ‘final’ boundaries were the most indicated by the annotators and, therefore, were considered the ideal segmentation for the text, as shown in the last row.

The final segmented text (translated from the original language, Portuguese) is shown in Figure 4. The main topic of the text is a plane crash, organised in three subtopics: the first text block is about the victims and where the plane was, the second block describes the plane, and the last one describes the crew.

[S1] A plane crash in Bukavu, in the Eastern Democratic Republic of Congo, killed 17 people on Thursday afternoon, said the spokesman of the United Nations.
[S2] The victims of the accident were 14 passengers and three crew members.
[S3] All died when the plane, hampered by the bad weather, failed to reach the runway and crashed in a forest that was 15 kilometers from the airport in Bukavu.
[S4] The plane exploded and caught fire, said the UN spokesman in Kinshasa, Jean-Tobias Okala.
[S5] "There were no survivors", said Okala.
<subtopic: plane crash in the Congo>

[S6] The spokesman said the plane, a Soviet Antonov-28 of Ukrainian manufacturing and ownership of the Trasept Congo, a Congolese company, also took a mineral load.
<subtopic: details about the plane>

[S7] According to airport sources, the crew members were Russian.
<subtopic: details about the flight crew>

[S1] Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo, matou 17 pessoas na quinta-feira à tarde, informou hoje um porta-voz das Nações Unidas.
[S2] As vítimas do acidente foram 14 passageiros e três membros da tripulação.
[S3] "Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 Km do aeroporto de Bukavu.
[S4] O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.
[S5] "Não houve sobreviventes", disse Okala.
<subtópico: estado atual>

[S6] O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.
<subtópico: detalhes do avião>

[S7] Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.
<subtópico: mensagens de apoio>

Figure 4: Example of a text with identified subtopics.

Figure 5 provides an example of where all the annotators placed boundaries in the same places. The main topic is Thiago Pereira's participation on the Pan-American Games (Thiago Pereira is a Brazilian swimmer). The text (translated from the original language, Portuguese) was segmented in three subtopics: the first one is Thiago Pereira, the second is about other athletes, and the third subtopic provides background information about Thiago Pereira. We believe that, the simpler the text content is, the greater the chance of having good agreement between annotators is.

Figure 6 shows the number of subtopics in the reference segmentation. There were eight texts with only one subtopic, twenty-four texts with two subtopics, fifty texts with three subtopics, thirty-one texts with four subtopics, nineteen texts with five subtopics, four texts with six subtopics, two texts with seven subtopics, and two texts with eight subtopics. Overall, the average number of subtopic boundaries in a text is three. Most of the boundaries (99 percent) occur between paragraphs. This is because writers usually structure their texts in a way that paragraphs constitute the basic thematic organisation (one paragraph, one subtopic). The descriptions given by each judge after a subtopic boundary were not used to define the final annotation. In this study, the descriptions were used only to better understand the annotators' decisions.

[S1] Thiago Pereira, a swimmer from Brazil, won the gold medal in 4x200 medley, the second for Brazil.
[S2] The Brazilian led the competition and hit the Pan American and South American record with a time of 4min11s14.
<subtopic: Thiago Pereira>

[S3] The American Robert Margalis won the silver medal.
[S4] The Canadian Keith Beavers won the bronze medal.
[S5] Another Brazilian, Diogo Yabe, stayed in fourth position.
<subtopic: other athletes>

[S6] Thiago Pereira had already won the silver medal in 2003 in the same competition of 4x200 medley, in Santo Domingo.
<subtopic: Thiago Pereira's history>

[S1] O brasileiro Thiago Pereira conquista com folga o primeiro lugar nos 4x200m medley, levando a medalha de ouro, segundo do Brasil.
[S2] O brasileiro liderou de ponta a ponta e ao final bateu o recorde Pan-Americano e Sul-Americano, com o tempo de 4min11s14.
<subtópico: Thiago Pereira>

[S3] O americano Robert Margalis ficou com a prata.
[S4] O canadense Keith Beavers garantiu o bronze.
[S5] O outro brasileiro, Diogo Yabe, ficou com a quarta colocação.
<subtópico: demais atletas>

[S6] Thiago Pereira, já havia sido bronze em 2003 na mesma prova de 4x200m medley, em Santo Domingo.
<subtópico: histórico de Thiago Pereira>

Figure 5: Example of text with full agreement on subtopics.

4.7 Delivering and maintaining the product

The corpus and its annotation are available for research purposes. For each text, we provide the reference annotation and all the segmentations performed by each annotator. We believe the data will be useful for researchers interested in investigating subtopic segmentation, but also for work that examines the influence of other text characteristics on each annotator's behaviour. The corpus is stored in plain text format, which we adopted due to its simplicity.

5. Strategies for subtopic segmentation

This section describes our proposal for automatically identifying and partitioning the subtopics of a text. We developed four baseline and six other algorithms that are based on discourse features.

The four baseline algorithms segment at paragraphs, sentences, random boundaries (randomly selecting any number of boundaries and where they are in a text), or are based on word reiteration. The word reiteration strategy is an adaptation of the well-known TextTiling method (Hearst, 1997) for the characteristics of the corpus that we used.[5] We did not test other

[5] More specifically, we have used the block comparison method with block size = 2.

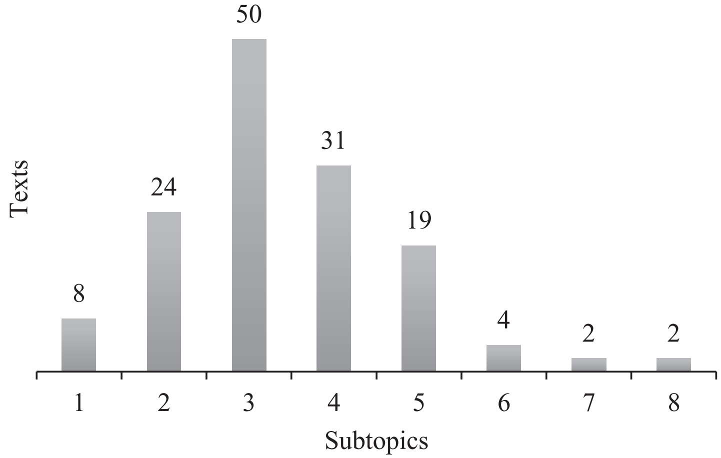

Figure 6: Number of subtopics in texts.

methods from the literature because our focus was on exploring the discourse structure influence in subtopic segmentation.

The algorithms based on discourse consider the discourse structure itself and the RST relations in the discourse tree. The first algorithm (which we refer to as Simple_Cosine) is based on Marcu (2000a) for measuring the 'goodness' of a discourse tree. The author assumes that a discourse tree is 'better' if it exhibits a high-level structure that matches as much as possible the subtopic boundaries of the text for which that structure was built. Marcu associates a clustering score to each node in a tree: for the leaves, this score is 0; and for the internal nodes, the score is given by the lexical similarity between the immediate children. The hypothesis underlying such measurements is that better trees show higher similarity among their nodes. We have adopted the same idea using the cosine measure (Salton, 1989) and proposed that text segments with similar vocabulary are likely to be part of the same subtopic segment. In our case, nodes with scores below the average score in the discourse tree are supposed to indicate possible subtopic boundaries.

The second algorithm, referred to as Cosine_Nuclei, is also a proposal by Marcu (2000a). It is assumed that, whenever a discourse relation holds between two textual spans, that relation also holds between the most salient units (nuclei) associated to those spans. We have used this formalisation and measured the similarity between the salient units associated to two spans (instead of measuring among all the text spans of the relation, as in the previous algorithm).

The third (Simple_Cosine_with_Depth) and fourth (Cosine_Nuclei_with_Depth) algorithms are variations of Simple_Cosine and Cosine_Nuclei, respectively. For these new strategies, the similarity for each node is divided by the depth where it occurs, traversing the tree in a

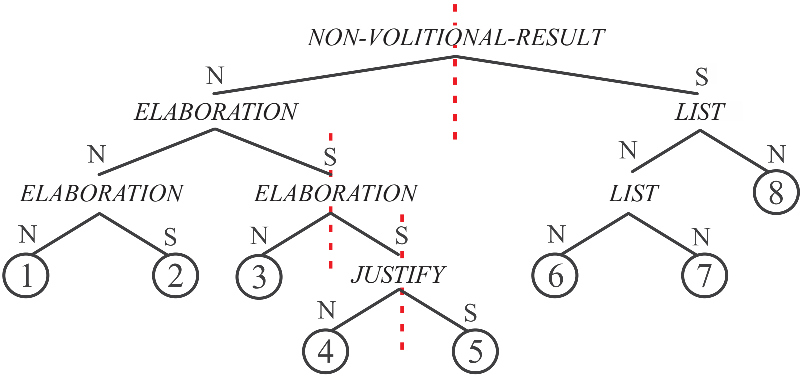

Figure 7: Example of an RST structure segmented in subtopics with Simple_Cosine_with_Depth.

bottom-up way. These should guarantee that higher nodes are 'weaker' and might better represent subtopic boundaries. Therefore, we assume that subtopic boundaries are more likely to be mirrored at the higher levels of the discourse structure. We have also used the average score to establish which are the less similar nodes.

Figure 7 shows the RST tree for the text presented in Figure 1. The dashed lines indicate the boundaries according to the Simple_Cosine_with_Depth segmentation method. For this tree, the similarity score between Nodes 4 and 5 is divided by 1 (since we are at the leaf level); the similarity score between Node 3 and ELABORATION is divided by 2 (since we are at a higher level, 1 above the leaves on the right); the score between ELABORATION and ELABORATION is divided by 3; and the score between ELABORATION and LIST is divided by 4. Comparing this automatic segmentation with the reference segmentation (presented in Figure 1), the Simple_Cosine_with_Depth method did not identify a boundary between Sentences 2 and 3 ('Elaboration' span), but it placed a correct boundary between Sentences 5 and 6 ('Non-Volitional-Result' span). The method also identified wrong boundaries between Sentences 3 and 4 ('Elaboration' span), and 4 and 5 ('Justify' span).

The next algorithms are based on the idea that some relations are more likely to represent subtopic shifts. For estimating this, we have used the CSTNews corpus. In this corpus, there are twenty-nine different types of RST relations that may connect sentences. Figure 8 shows those relations and their frequency among sentences. In general, ELABORATION is very common in diverse corpora in different languages; for example, in the RST Spanish Treebank (da Cunha *et al*., 2011), in Discourse Treebank (Carlson *et al*., 2003) or CorpusTCC (Pardo and Nunes, 2004). That is because Elaboration is a common rhetorical strategy which the writer may use to expand on the previous context – thus, it becomes a *de facto* default whenever a more semantically marked relation does not fit the context (Carlson *et al*., 2003).

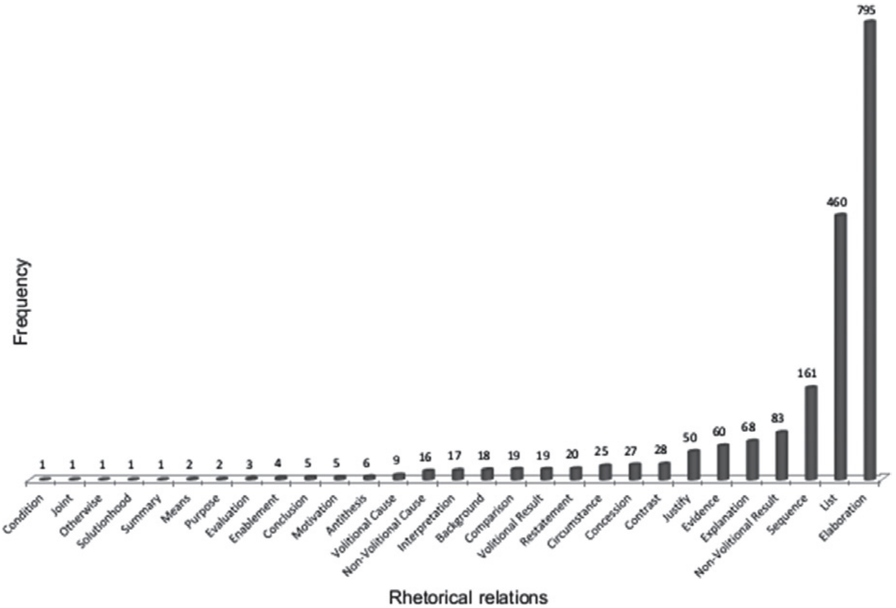

Figure 8: Number of RST relations in the CSTNews corpus.

The same occurs in CSTNews: ELABORATION relations are more frequent than others such as CONDITION and JOINT.

We also recorded the frequency of those relations in subtopic boundaries. We realised that some relations were more frequent in boundaries, whereas others never occurred at boundary points. Out of the twenty-nine relations, sixteen appeared in the reference annotation. Although the total inventory of relations is quite extensive, Figure 9 shows that, in the subtopic boundaries, ELABORATION was the most frequent relation (appearing in 57 percent of the boundaries), followed by LIST (19 percent) and NON-VOLITIONAL RESULT (5 percent). SEQUENCE and EVIDENCE appeared in 2 percent of the subtopic boundaries, and BACKGROUND, CIRCUMSTANCE, COMPARISON, CONCESSION, CONTRAST, EXPLANATION, INTERPRETATION, JUSTIFY, NON-VOLITIONAL CAUSE, VOLITIONAL CAUSE and VOLITIONAL RESULT in 1 percent of the boundaries.

Although there is a significant difference between the RST relations present in the corpus and those in subtopic boundaries, we searched for a way to use this information about frequencies to segment texts. In the literature, there are some works that assign weights for relations based on some classification. For instance, O'Donnell (1997) and Uzêda *et al.* (2010) developed methods that assume that each RST relation has an associated relevance score that indicates how important the respective segments are for the summary. In the same way, we took advantage of the relations' frequency on subtopic boundaries in the reference corpus, as well as their definitions

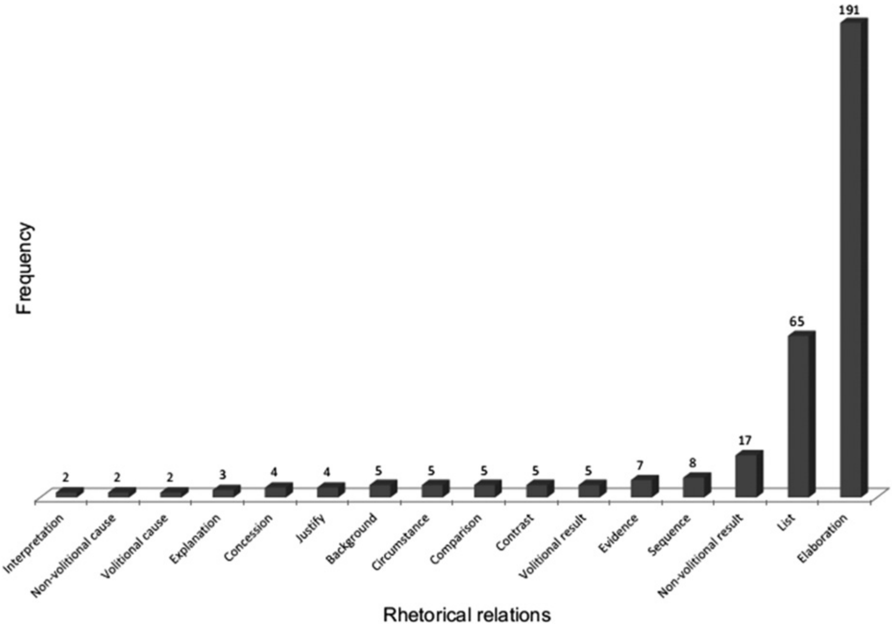

Figure 9: Number of occurrences of RST relations in subtopics boundaries of the CSTNews corpus.

Class	Relations
Weak (0.4)	Elaboration, Contrast, Joint, List
Medium (0.6)	Antithesis, Comparison, Evaluation, Means, Non-Volitional Cause, Non-Volitional Result, Solutionhood, Volitional Cause, Volitional Result, Sequence
Strong (0.8)	Background, Circumstance, Concession, Conclusion, Condition, Enablement, Evidence, Explanation, Interpretation, Justify, Motivation, Otherwise, Purpose, Restatement, Summary

Table 3: Classification of RST relations.

and our assumptions, to attribute a weight associated with the possibility that a relation indicates a boundary.

Table 3 shows how the twenty-nine relations were distributed. One relation is weak if it usually indicates a boundary; in this case, its weight is 0.4. One relation is medium because it may indicate a boundary or not; therefore, its weight is 0.6. On the other hand, a strong relation almost never indicates a subtopic boundary; therefore, its weight is 0.8. Such weights were empirically determined. We do not allow relations to have a weight of 1,

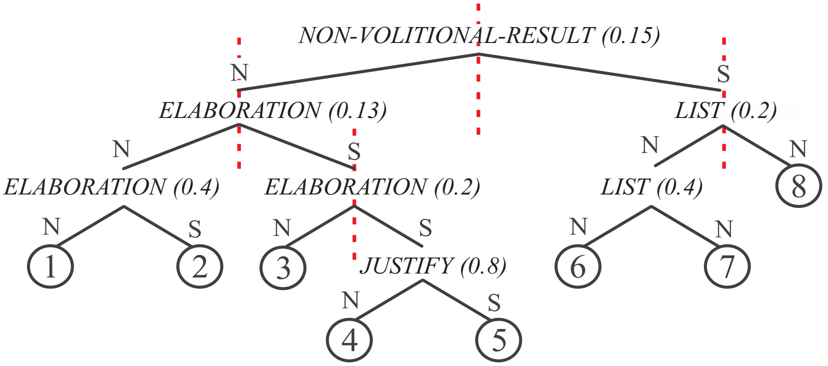

Figure 10: Example of an RST structure segmented in subtopics with Relation_with_Depth.

because we consider that it is appropriate that, whenever a relation is found (going up in the tree), the probability of finding a boundary increases.

A factor that may be observed is that all presentational relations are classified as strong, with the exception of antithesis. This is related to the definition of presentational relations, which are used to increase some inclination on the part of the reader, and are not, therefore, more likely to begin or end a subtopic. ANTITHESIS is an exception because it was found in the reference segmentation with low frequency.

From this classification, we created two more strategies: Relation_with_Depth and Cosine_Nuclei_with_Depth_Relation. The strategy Relation_with_Depth associates a score to the nodes by dividing the relation's weight by the depth where it occurs, in a bottom-up way of traversing the tree. We have also used the average score to find nodes that are less similar. As we have observed that some improvement might be achieved every time nucleus information was used, we tried to combine this configuration with the relations' weight. Hence, we computed the scores of the Cosine_Nuclei_with_Depth strategy times the proposed relation's weight. We named this algorithm Cosine_Nuclei_with_Depth_Relation.

Figure 10 shows the segmentation produced using the strategy Relation_with_Depth for the text presented in Figure 1. The numbers in brackets represent the score at each level. The average among nodes scores is 0.3; in this case, Relation_with_Depth made two correct boundaries: between Sentences 2 and 3, and 5 and 6. However, there two wrong automatic boundaries: between Sentences 3 and 4, and 7 and 8.

6. Results and discussion

There are several ways to evaluate a segmentation algorithm, including comparing its segmentation against that of human judges and comparing its segmentation against other automated segmentation strategies. This section presents comparisons of the results of the algorithms over the reference corpus. The performance of subtopic segmentation is usually measured using Recall (R), Precision (P) and F-measure (F) scores. These scores quantify how closely the system subtopics correspond to the ones produced by humans. Recall (R) is the percentage of proposed boundaries that exactly match boundaries in the reference segmentation. Precision (P) is the percentage of reference segmentation boundaries that are identified by the algorithm. F-measure (F) is defined as the harmonic mean of P and R, being a unique measure of overall system performance.

Those measures compare the boundary correspondences without considering whether these are close to each other: if they are not the same (regardless of whether they are closer to or farther away from one another), they score zero. However, it is also important to know how close the identified boundaries are to the expected ones, since this may help to determine how serious the errors made by the algorithms are.

Pevzner and Hearst (2002) propose an alternative metric called WindowDiff (WD) for subtopic segmentation. WD works as follows: for each interval (k), simply compare the number of reference segmentation boundaries that fall in this interval (Ri) with the number of boundaries that are assigned by the algorithm (Ai). The algorithm is penalised if Ri $\neq$ Ai (which is computed as |Ri – Ai| > 0). Equation 1 shows the definition of WindowDiff, where b(i, i+k) represents the number of boundaries between positions *i* and *i+k* in the text, and *N* is the total number of sentences in the text. The smaller the WD value of a segmentation method is, the better its performance in detecting subtopic boundaries is. Equation 2 represents the window size (k).

$$WindowDiff = \frac{1}{N-K}\sum\nolimits_{i=1}^{N-k}(|b(r_i, r_{i+k}) - b(a_i, a_{i+k})| > 0) \quad (1)$$

$$K = \frac{N}{2 * number\ of\ segments} \quad (2)$$

We also propose a simpler measure to this, which we call Deviation (D) from the reference annotation. Considering two algorithms that propose the same number of boundaries for a text and make one single mistake each (having, therefore, the same P, R and F scores), the best one will be the one that deviates the least from the reference.

Figure 11 provides one example of how to compute the D measure. Consider a hypothetical text with nine sentences (represented by the rectangles), where ‘reference’ indicates the manual segmentation and ‘automatic’ is the same text with automatically determined boundaries. Boundaries are indicated by vertical lines among sentences. For each automatic boundary, we compute its distance in relation to the closest

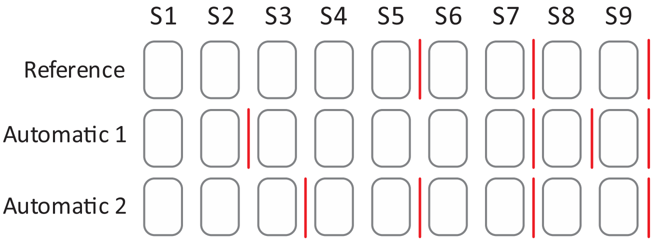

Figure 11: Example for Deviation measure.

reference boundary. One may notice that the last boundary of the reference is also present on the automatic segmentation. In this case, the difference between them is zero (D = 0, so far). However, the 'Automatic 1' algorithm missed the boundary between Sentences 5 and 6 (false negative), and placed two boundaries that do not exist in the reference segmentation (false positive), between Sentences 2 and 3, and Sentences 8 and 9. The distance between automatic segmentation (between Sentences 2 and 3) with respect to the nearest reference segmentation is 3. The automatic segmentation between Sentences 8 and 9 is one sentence away from the nearest reference segmentation. Therefore, the overall deviation given by the sum of all deviations is 4. The 'Automatic 2' algorithm placed one boundary between Sentences 3 and 4 (false positive); and it has D = 2, consequently. Finally, we normalise the values according to the highest D across the algorithms, and find the final D for each one. In this case, 'Automatic 1' has D = 1 and 'Automatic 2' has D = 0.5. The lower the D, the better the result is.

The difference between WD and D is that the first compares two segments and, if they are different, the algorithm is penalised, no matter whether the automatic boundary is far from the reference boundary or not. D, on the other hand, specifies the shortest distance between an automatic and a reference boundary. Both measures produce values between 0 and 1, with 1 being worst. Such measures aim to smooth the results of R, P and F, which heavily penalise an algorithm that does not segment the text in exactly the same places as the reference segmentation. It is important to say that all these metrics have their own problems. It is hard to know exactly what each of them intuitively means in isolation; it is necessary to have other systems' performance values for purposes of comparison (Purver, 2011).

Table 4 shows the results of the methods we investigated. The first four rows represent the baselines and the six following rows, the proposed algorithms based on RST. The last row shows the human performance, which we refer to as topline. A topline indicates the upper bound results that automatic methods may achieve in the task. To find the topline, a human annotator of the corpus was randomly selected for each text and its annotation was compared to the reference text.

As expected, the paragraph baseline was very good, having excellent R and F values in the baseline set. This shows that, in most of the texts, the

Algorithm	R	P	F	WD	D
TextTiling	0.405	0.773	0.497	0.375	0.042
Paragraph	0.989	0.471	0.613	0.591	0.453
Sentence	1.000	0.270	0.415	0.892	1.000
Random	0.674	0.340	0.416	0.669	0.539
Simple_Cosine	0.549	0.271	0.345	0.694	0.545
Cosine_Nuclei	0.631	0.290	0.379	0.691	0.556
Simple_Cosine_with_Depth	0.873	0.364	0.489	0.711	0.577
Cosine_Nuclei_with_Depth	0.899	0.370	0.495	0.710	0.586
Relation_with_Depth	0.901	0.507	0.616	0.525	0.335
Cosine_Nuclei_with_Depth_Relation	0.908	0.353	0.484	0.729	0.626
Topline (human)	0.807	0.799	0.767	0.182	0.304

Table 4: Evaluation of algorithms.

subtopics are organised in paragraphs. Although the Sentence baseline has the best R, it has the worst P, WD and D. This is due to the fact that not every sentence is a subtopic, and segmenting all of them becomes a problem when we are looking for major groups of subtopics. Random also has bad values for F, WD and D. TextTiling is the algorithm that deviates the least from the reference segmentation (WD and D values). This happens because it is very conservative and detects only a few segments, sometimes only one (the end of the text), causing it to have good P, WD and D scores, but heavily penalising R.

In the case of the algorithms based on RST, we may notice that some of them produced good results in terms of R, P and F, with acceptable WD and D values. We note too that every time the salient units were used, R and P increase, except for Cosine_Nuclei_with_Depth_Relation. Examining the measures, we notice that, among the algorithms based on RST, the best one was Relation_with_Depth. Although its F is close to the one of the Paragraph baseline, the Relation_with_Depth algorithm shows much better D and WD values. One may see that the Relation_with_Depth algorithm and the traditional TextTiling have the best WD and D values.

As expected, the Topline (the human, therefore) achieved the best F with acceptable D and WD. Its F value is probably the best that an automatic method may expect to achieve in our corpus. It is 25 percent better than our best method (Relation_with_Depth). There is, therefore, room for improvement, possibly using other discourse features, such as discourse markers.

Table 5 shows the average number of segments predicted by each algorithm per text. TextTiling segments less than all the other algorithms. This is related to text sizes, which, in their majority, are not very long.

Algorithm	Average boundaries
TextTiling	1.6
Paragraph	8.1
Sentence	14.8
Random	8.0
Simple_Cosine	8.1
Cosine_Nuclei	8.5
Simple_Cosine_with_Depth	9.5
Cosine_Nuclei_with_Depth	9.8
Relation_with_Depth	6.9
Cosine_Nuclei_with_Depth_Relation	10.15
Topline (human)	3

Table 5: Average segments per algorithm.

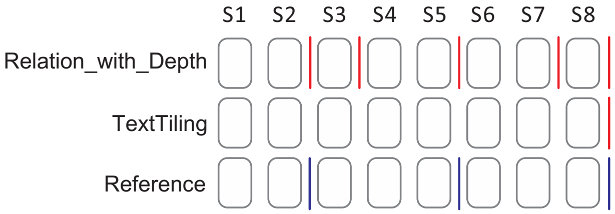

Figure 12: Comparison between TextTiling and Relation_with_Depth.

TextTiling tends not to perform well on short texts (although we have adapted it to the text genre/type). One may also notice that Relation_with_Depth does not segment as much as other algorithms based on RST. Therefore, considering the discourse features, Relation_with_Depth shows the best cost–benefit relation: the best R and F results, acceptable WD and D, and a slightly more conservative number of segments (which is, somehow, not surprising for news texts, as we have already shown that such texts have on average three segments).

We have run t-tests for pairs of algorithms for which we wanted to check the statistical difference. As expected, the F difference is not significant for Relation_with_Depth and the Paragraph algorithms, but it was significant with 95 percent confidence for the comparison of Relation_with_Depth and TextTiling (also regarding the F values). Finally, the difference between Relation_with_Depth and the Topline was also significant. The same occurs for WD and D comparisons.

The results demonstrate that discourse organisation mirrors subtopic changes in the texts. Although the six algorithms based on discourse features have excellent R, they still segment too much. Figure 12 compares different segmentations for the text presented in Figure 1, made by TextTiling and

Relation_with_Depth (the segmented tree is presented in Figure 10), which show the best results. While TextTiling did not find boundaries in the text, Relation_with_Depth found two true boundaries (between Sentences S2 and S3, and S5 and S6) and two near-misses (between sentences S3 and S4, and S7 and S8). As in other examples of segmentation made by Relation_with_Depth, these two near-misses are at lower levels of tree. In this case, TextTiling has F = 0.5, WD = 0.25 and D = 0; Relation_with_Depth has F = 0.75, WD = 0.25 and D = 0.4. As we assume that subtopic boundaries are more likely to be mirrored at the higher levels of the discourse structure, Relation_with_Depth needs some adjustment in order not to segment so much at lower levels of the tree.

The methods for subtopic segmentation are language independent, but, for other genres, it is possible that some adjustment may be necessary. The rules for subtopic segmentation were extracted from news texts. For scientific texts, for instance, the rules may not work well.

In order to apply the methods over non-annotated texts, discourse parsers exist for English (Marcu, 2000a), Spanish (Maziero *et al.*, 2011) and Portuguese (Pardo and Nunes, 2008), and might be used.

7. Conclusion

We have presented the main questions regarding corpus annotation for the phenomenon of subtopic annotation and described several algorithms for subtopic segmentation based on discourse features.

With regard to the subtopic annotation, our main contributions are two-fold: discussing and performing the annotation process in a systematic way, and making available a valuable reference corpus for subtopic study. The corpus does not only contain the reported annotations, but several other annotations, and information, that are useful for several NLP tasks. It also includes single and multi-document discourse annotation, text-summary alignments, different types of summaries for each text and cluster, temporal and aspect annotations, and word sense annotation for nouns and verbs, among others. As described before, the labels for each subtopic are available for each manually annotated text. Such labels were mentioned but not actually used in this work. In the future, we plan to investigate how to produce labels for subtopics similar to the ones indicated in the manual annotation.

We proposed different algorithms for subtopic segmentation, which were evaluated over the reference corpus. The results demonstrate that discourse knowledge can help find boundaries in a text. In particular, the relation type and the level in the discourse structure in which the relation happens are important features. To the best of our knowledge, this is the first attempt to correlate RST structures with subtopic boundaries, which we believe is an important theoretical advance. As for subtopic segmentation, the labels of subtopics, assigned by annotators, could also be exploited in order to improve the segmentation algorithms. The labels may be combined with rhetorical and semantic information (such as lexical chain structure) in order to improve segmentation.

Acknowledgments

The authors are grateful to FAPESP, CAPES, CNPq and Natural Sciences and Engineering Research Council of Canada (Discovery Grant 261104–2008) for supporting this work.

References

Aluísio, S.M., T.A.S. Pardo and M.S. Duran. 2014. 'New corpora for new challenges in Portuguese processing' in T.B. Sardinha and T.L.S.B. Ferreira (eds) Working with Portuguese Corpora, pp. 303–22. London and New York: Bloomsbury Academic.

Blei, D.M., A.Y. Ng and M.I. Jordan. 2003. 'Latent dirichlet allocation', Journal of Machine Learning Research, Vol. 3, pp. 993–1022.

Bollegala, D., N. Okazaki and M. Ishizuka. 2006. 'A bottom-up approach to sentence ordering for multi-document summarization', Journal Information Processing and Management 46 (1), pp. 89–109.

Burger, S., V. MacLaren and H. Yu. 2002. 'The ISL meeting corpus: the impact of meeting type on speech style', in Proceedings of the International Conference Spoken Language Processing (2002), pp. 1–4. 16–20 September. Denver, USA.

Cardoso, P.C.F., E.G. Maziero, M.L.C. Jorge, E.M.R. Seno, A. Di Felippo, L.H.M. Rino, M.G.V. Nunes and T.A.S. Pardo. 2011. '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, pp. 88–105. 26 October. Cuiabá/MT, Brazil.

Carletta, J. 1996. 'Assessing agreement on classification tasks: the Kappa statistic', Computational Linguistics 22 (2), pp. 249–54.

Carlson, L. and D. Marcu. 2001. 'Discourse tagging reference manual', Technical Report ISI-TR-545, University of Southern, California.

Carlson, L., D. Marcu and M.E. Okurowski. 2003. 'Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory' in Proceedings of 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001. Aalborg, Denmark.

Chang, T.H. and C.H. Lee. 2003. 'Topic segmentation for short texts' in Proceedings of the 17th Pacific Asia Conference Language, pp. 159–65. 24–26 June. Sentosa, Singapore.

Choi, F.Y.Y. 2000. 'Advances in domain independent linear text segmentation' in Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2000), pp. 26–36. 29 April to 4 May. Seattle, Washington.

da Cunha, I., J.-M. Torres-Moreno and G. Sierra. 2011. 'On the development of the RST Spanish Treebank' in Proceedings of the 5th Linguistic Annotation Workshop (LAW-2011), pp. 1–10. 23–29 June. Portland-Oregon.

Du, L., W.L. Buntine and M. Johnson. 2013. 'Topic Segmentation with a Structured Topic Model' in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 190–200. 10–15 June. Atlanta, Georgia.

Francis, W.N. and H. Kučera. 1979. Brown Corpus Manual. Linguistics Department, Brown University.

Galley, M., K. Mckeown, E. Fosler-Lussier and H. Jing. 2003. 'Discourse segmentation of multi-party conversation' in Proceedings of the 41st Annual Conference of the Association for Computational Linguistics (ACL-2003), pp. 562–9. 7–12 July. Sapporo, Japan.

Gruenstein, A., J. Niekrasz and M. Purver. 2007. 'Meeting structure annotation: annotations collected with a general purpose toolkit' in L. Dybkjaer and W. Minker (eds) Recent Trends in Discourse and Dialogue, pp. 247–74. Dordrecht: Springer.

Hearst, M. 1993. 'TextTiling: a quantitative approach to discourse segmentation'. (Technical report.) University of California, Berkeley, Sequoia.

Hearst, M. 1994. 'Multi-paragraph segmentation of expository text' in Proceedings of the 32nd annual meeting on Association for Computational Linguistics (ACL-1994). 27–30 June. Las Cruces, New Mexico.

Hearst, M. 1997. 'TextTiling: segmenting text into multi-paragraph subtopic passages', Computational Linguistics 23 (1), pp. 33–64.

Hennig, L. 2009. 'Topic-based multi-document summarization with probabilistic latent semantic analysis', Recent Advances in Natural Language Processing – RANLP (2009), pp. 144–9. Borovets, Bulgaria.

Hovy, E. 2009. 'Text summarization' in The Oxford Handbook of Computational Linguistics, pp. 583–98. United States: Oxford University.

Hovy, E. and J. Lavid. 2010. 'Towards a science of corpus annotation: a new methodological challenge for Corpus Linguistics', International Journal of Translation Studies 22 (1), pp. 13–36.

Hovy, E. and C.-Y. Lin. 1998. 'Automated text summarization and the SUMMARIST system' in Proceedings of a Workshop on Held at Baltimore, pp. 197–214. 13–15 October. Baltimore, Maryland.

Iruskieta, M., M.J. Aranzabe, A.D. Ilarraza, I. Gonzalez-Dios, M. Lersundi and O.L. Lacalle. 2013. 'The RST Basque Treebank: an online search interface to check rhetorical relations' in Proceedings of 4th Workshop RST and Discourse Studies, pp. 40–9. 24–26 October. Fortaleza/CE, Brazil.

Janin, A., D. Baron, D.E. Edwards, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke and C. Wooters. 2003. 'The ICSI meeting corpus' in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 364–7. 6–10 April. Hong Kong, China.

Kazantseva, A. and S. Szpakowicz. 2012. 'Topical segmentation: a study of human performance and a new measure of quality' in Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 211–20. 3–8 June. Montreal, Canada.

Knott, A., J. Oberlander, M. O'Donnel and C. Mellish. 2001. 'Beyond elaboration: the interaction of relations and focus in coherent text' in T. Sanders, J. Schilperoord and W. Spooren (eds) Text Representation: Linguistic and Psycholinguistic Aspects, pp. 181–96. Amsterdam: John Benjamins.

Koch, I.G.V. 2009. Introdução à linguística textual. São Paulo: Contexto.

Lage, N. 2002. Estrutura da Notícia. (Fifth edition.) São Paulo: Ática.

Leech, G. 2005. 'Adding linguistic annotation' in M. Wynne (ed.) Developing Linguistic Corpora: A Guide to Good Practice, pp. 25–38. Oxford: Oxbow Books.

Mann, W.C. and S.A. Thompson. 1987. 'Rhetorical Structure Theory: a theory of text organization', Technical Report ISI/RS-87–190. Marina del Rey, California: Information Sciences Institute.

Mann, W.C. and S.A. Thompson. 1988. 'Rhetorical Structure Theory: toward a functional theory of text organization', Text 8 (3), pp. 243–81.

Marcu, D. 2000a. The Theory and Practice of Discourse Parsing and Summarization. Cambridge, Massachusetts: MIT Press.

Marcu, D. 2000b. 'The rhetorical parsing of unrestricted texts: a surface-based approach', Computational Linguistics 26 (3), pp. 395–448.

Maziero, E.G. and T.A.S. Pardo. 2009. 'Automatização de um Método de Avaliação de Estruturas Retóricas' in Proceedings of the 3rd RST Brazilian Meeting, pp. 1–9. 8–11 September. São Carlos-SP, Brazil.

Maziero, E.G., T.A.S. Pardo, I. da Cunha, J.M. Torres-Moreno and E. SanJuan. 2011. 'DiZer 2.0 – an adaptable on-line discourse parser' in Proceedings of the 3rd RST Brazilian Meeting, pp. 1–17. 26 October. Cuiabá/MT, Brazil.

O'Donnell, M. 1997. 'Variable-length on-line document generation' in Proceedings of the 6th European Workshop on Natural Language Generation, pp. 1–5. 24–26 March. Duiburg, Germany.

Oh, H.-J., S.H. Myaeng and M.-G. Jang. 2007. 'Semantic passage on sentence topics for question answering', Information Sciences 177 (18), pp. 3696–717.

Pardo, T.A.S. and M.G.V. Nunes. 2004. 'Relações Retóricas e seus Marcadores Superficiais: Análise de um Corpus de Textos Científicos em Português do Brasil'. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, N. 231. São Carlos-SP, 73p.

Pardo, T.A.S. and M.G.V. Nunes. 2008. 'On the development and evaluation of a Brazilian Portuguese discourse parser', Journal of Theoretical and Applied Computing 15 (2), pp. 43–64.

Passonneau, R.J. and D.J. Litman. 1997. 'Discourse segmentation by human and automated means', Computational Linguistics 23 (1), pp. 103–9.

Pevzner, L. and M. Hearst. 2002. 'A critique and improvement of an evaluation metric for text segmentation', Computational Linguistics 28 (1), pp. 19–36.

Prince, V. and A. Labadié. 2007. 'Text segmentation based on document understanding for information retrieval' in Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems, pp. 295–304. 27–29 June. Paris, France.

Purver, M. 2011. 'Topic segmentation' in G. Tur and R. de Mori (eds) Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, pp. 291–317. Hoboken, New Jersey: Wiley.

Riedl, M. and C. Biemann. 2012. 'TopicTiling: a text segmentation algorithm based on LDA' in Proceedings of the 2012 Student Research Workshop, Association for Computational Linguistics, pp. 37–42. 3–8 June. Jeju, Republic of Korea.

Salton, G. 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of. Addison-Wesley.

Taboada, M. and W.C. Mann. 2006. 'Rhetorical structure theory: looking back and moving ahead', Discourse Studies 8 (3), pp. 423–59.

Taboada, M. and D. Das. 2013. 'Annotation upon annotation: adding signaling information to a corpus of discourse relations', Dialogue and Discourse 4 (2), pp. 249–81.

Uzêda, V.R., T.A.S. Pardo and M.G.V. Nunes. 2010. ‘A comprehensive comparative evaluation of RST-based summarization methods’, ACM Transactions on Speech and Language Processing 6 (4), pp. 1–20.

Wan, X. 2008. ‘An exploration of document impact on graph-based multi-document summarization’ in Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 755–62. 25–27 October. Waikiki, Honolulu, Hawai’i.

Your short guide to the EUP Journals Blog http://euppublishingblog.com/

A forum for discussions relating to Edinburgh University Press Journals

1. The primary goal of the EUP Journals Blog

To aid discovery of authors, articles, research, multimedia and reviews published in Journals, and as a consequence contribute to increasing traffic, usage and citations of journal content.

2. Audience

Blog posts are written for an educated, popular and academic audience within EUP Journals' publishing fields.

3. Content criteria - your ideas for posts

We prioritize posts that will feature highly in search rankings, that are shareable and that will drive readers to your article on the EUP site.

4. Word count, style, and formatting

- Flexible length, however typical posts range 70-600 words.
- Related images and media files are encouraged.
- No heavy restrictions to the style or format of the post, but it should best reflect the content and topic discussed.

5. Linking policy

- Links to external blogs and websites that are related to the author, subject matter and to EUP publishing fields are encouraged, e.g.to related blog posts

6. Submit your post

Submit to ruth.allison@eup.ed.ac.uk

If you'd like to be a regular contributor, then we can set you up as an author so you can create, edit, publish, and delete your *own* posts, as well as upload files and images.

7. Republishing/repurposing

Posts may be re-used and re-purposed on other websites and blogs, but a minimum 2 week waiting period is suggested, and an acknowledgement and link to the original post on the EUP blog is requested.

8. Items to accompany post

- A short biography (ideally 25 words or less, but up to 40 words)
- A photo/headshot image of the author(s) if possible.
- Any relevant, thematic images or accompanying media (podcasts, video, graphics and photographs), provided copyright and permission to republish has been obtained.
- Files should be high resolution and a maximum of 1GB
- Permitted file types: *jpg, jpeg, png, gif, pdf, doc, ppt, odt, pptx, docx, pps, ppsx, xls, xlsx, key, mp3, m4a, wav, ogg, zip, ogv, mp4, m4v, mov, wmv, avi, mpg, 3gp, 3g2***.**