Cross-lingual and cross-domain discourse segmentation of entire documents

Chloé Braud, Ophélie Lacroix, Anders Søgaard

HAL Id: hal-02373807
https://hal.science/hal-02373807

Cross-lingual and cross-domain discourse segmentation
of entire documents

Chloé Braud
CoAStaL DIKU
University of Copenhagen
University Park 5,
2100 Copenhagen
chloe.braud@gmail.com

Ophélie Lacroix
CoAStaL DIKU
University of Copenhagen
University Park 5,
2100 Copenhagen
lacroix@di.ku.dk

Anders Søgaard
CoAStaL DIKU
University of Copenhagen
University Park 5,
2100 Copenhagen
soegaard@di.ku.dk

Abstract

Discourse segmentation is a crucial step in building end-to-end discourse parsers. However, discourse segmenters only exist for a few languages and domains. Typically they only detect intra-sentential segment boundaries, assuming gold standard sentence and token segmentation, and relying on high-quality syntactic parses and rich heuristics that are not generally available across languages and domains. In this paper, we propose statistical discourse segmenters for five languages and three domains that do not rely on gold pre-annotations. We also consider the problem of learning discourse segmenters when no labeled data is available for a language. Our fully supervised system obtains 89.5% F_1 for English newswire, with slight drops in performance on other domains, and we report supervised and unsupervised (cross-lingual) results for five languages in total.

1 Introduction

Discourse segmentation is the first step in building a discourse parser. The goal is to identify the minimal units — called Elementary Discourse Units (EDU) — in the documents that will then be linked by discourse relations. For example, the sentences (1a) and (1b)[1] are each segmented into two EDUs, then respectively linked by a CONTRAST and an ATTRIBUTION relation. The EDUs are mostly clauses and may cover a full sentence. This step is crucial: making a segmentation error leads to an error in the final analysis. Discourse segmentation can also inform other tasks, such as argumentation

mining, anaphora resolution, or speech act assignment (Sidarenka et al., 2015).

(1) a. [Such trappings suggest a glorious past] [but give no hint of a troubled present.]

b. [He said] [the thrift will to get regulators to reverse the decision.]

We focus on the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) – and resources such as the RST Discourse Treebank (RST-DT) (Carlson et al., 2001) – in which discourse structures are trees covering the documents. Most recent works on RST discourse parsing focuses on the task of tree building, relying on a gold discourse segmentation (Ji and Eisenstein, 2014; Feng and Hirst, 2014; Li et al., 2014; Joty et al., 2013). However, discourse parsers' performance drops by 12-14% when relying on *predicted* segmentation (Joty et al., 2015), underscoring the importance of discourse segmentation. State-of-the-art performance for discourse segmentation on the RST-DT is about 91% in F_1 with predicted parses (Xuan Bach et al., 2012), but these systems rely on a gold segmentation of sentences and words, therefore probably overestimating performance *in the wild*. We propose to build discourse segmenters without making any data assumptions. Specifically, rather than segmenting sentences, our systems segment documents directly.

Furthermore, only a few systems have been developed for languages other than English and domains other than the Wall Street Journal texts from the RST-DT. We are the first to perform experiments across 5 languages, and 3 non-newswire English domains. Since our goal is to provide a system usable for low-resource languages, we only use language-independent resources: here, the Universal Dependencies (UD) (Nivre et al.,

---

[1]The examples come from the RST Discourse Treebank.

2016) Part-of-Speech (POS) tags, for which annotations exist for about 50 languages. For the cross-lingual experiments, we also rely on cross-lingual word embeddings induced from parallel data. With a shared representation, we can transfer model parameters across languages, or learn models jointly through multi-task learning.

Contributions: We (i) propose a general statistical discourse segmenter (ii) that does not assume gold sentences and tokens, and (iii) evaluate it across 5 languages and 3 domains.

We make our code available at https://bitbucket.org/chloebt/discourse.

2 Related work

For English RST-DT, the best discourse segmentation results were presented in Xuan Bach et al. (2012) (F_1 91.0% with automatic parse, 93.7 with gold parse) – and in Joty et al. (2015) for the Instructional corpus (Subba and Di Eugenio, 2009) (F_1 80.9% on 10-fold). Segmenters based on handwritten rules have been developed for Brazilian Portuguese (Pardo and Nunes, 2008) (51.3% to 56.8%, depending on the genre), Spanish (da Cunha et al., 2010, 2012) (80%) and Dutch (van der Vliet, 2010) (73% with automatic parse, 82% with gold parse).[2]

Most statistical discourse segmenters are based on classifiers (Fisher and Roark, 2007; Joty et al., 2015). Subba and Di Eugenio (2007) were the first to use a neural network, and Sporleder and Lapata (2005) to model the task as a sequence prediction problem. In this work, we do sequence prediction using a neural network.

All these systems rely on a quite large range of lexical and syntactic features (e.g. token, POS tags, lexicalized production rules). Sporleder and Lapata (2005) present arguments for a knowledge-lean system that can be used for low-resourced languages. Their system, however, still relies on several tools and gold annotations (e.g. POS tagger, chunker, list of connectives, gold sentences). In contrast, we present what is to the best of our knowledge the first work on discourse segmentation that is directly applicable to low-resource languages, presenting results for scenarios where no labeled data is available for the target language.

Previous work, relying on gold sentence boundaries, also only considers intra-sentential segment

boundaries. We move to processing entire documents, motivated by the fact that sentence boundaries are not easily detected across all languages.

3 Discourse segmentation

Nature of the EDUs Discourse segmentation is the first step in annotating a discourse corpus. The annotation guidelines define what is the nature of the EDUs, broadly relying on lexical and syntactic clues. If sentences and independent clauses are always minimal units, some fine distinctions make the task difficult.

In the English RST-DT (Carlson and Marcu, 2001), lexical information is crucial: for instance, the presence of the discourse connective "but" in example (1a)[3] indicates the beginning of an EDU. In addition, clausal complements of verbs are generally not treated as EDUs. Exceptions are the complements of attribution verbs, as in (1b), and the infinitival clauses marking a PURPOSE relation as the second EDU in (2a). Note that, in this latter example, the first infinitival clause ("to cover up . . .") is, however, not considered as an EDU. This fine distinction corresponds to one of the main difficulties of the task. Another one is linked to coordination: coordinated clauses are generally segmented as in (2b), but not coordinated verb phrases as in (2c).

(2) a. [A grand jury has been investigating whether officials at Southern Co. accounting conspired **to** cover up their accounting for spare parts] [**to** evade federal income taxes.]

b. [they parcel out money] [so that their clients can find temporary living quarters,] [buy food] (. . .) [and replaster walls.]

c. [Under Superfund, those] [who owned, generated or transported hazardous waste] [are liable for its cleanup, (. . .)]

Finally, in a multi-lingual and multi-domain setting, note that all the corpora do not follow the same rules: for example, the relation ATTRIBUTION is only annotated in the English RST-DT and the corpora for Brazilian Portuguese, consequently, complements of attribution verbs are not segmented in the other corpora.

---

[2] For German (Sidarenka et al., 2015) propose a segmenter in clauses (that may be EDU or not).

[3] All the examples given come from (Carlson et al., 2001).

Binary task As in previous studies, we view segmentation as a binary task at the word level: a word is either an EDU boundary (label B, beginning an EDU) or not (label I, inside an EDU). This design choice is motivated by the fact that, in RST corpora, the EDUs cover the documents entirely, and that EDUs mostly are adjacent spans of text. An exception is when embedded EDUs break up another EDU, as in Example (3). The units 1 and 3 form in fact one EDU. We follow previous work on treating this as three segments, but note that this may not be the optimal solution.

(3) [But maintaining the key components (...)]₁
[− a stable exchange rate and high levels of imports −]₂
[will consume enormous amounts (...).]₃

Document-level segmentation Contrary to previous studies, we do not assume gold sentences: Since sentence boundaries are EDU boundaries, our system jointly predicts sentence and intra-sentential EDU boundaries.

4 Cross-lingual/-domain segmentation

Data is scarce for discourse. In order to build statistical segmenters for new, low-resourced languages and domains, we propose to combine corpora within a multi-task learning setting (Section 5) leveraging data from well-resourced languages or domains. Models are trained on several (source) languages (resp. domains) – each viewed as an auxiliary task – for building a system for a (target) language (resp. domain).

Cross-domain For cross-domain experiments, the models are trained on all the other (source) domains and parameters are tuned on data for the target domain. This allows us to improve performance when only few data points (i.e. development set) are annotated for a specific domain (semi-supervised setting).

Cross-lingual For cross-lingual experiments, we tune our system's parameters by training a system on the data for three languages with sufficient amounts of data (namely, German, Spanish and Brazilian Portuguese), and using English data as a development set. We then train a new model also using multi-task learning (with these tuned parameters) using only source training data, and report performance on the target test set. This allows us to estimate performance when no data is available for the language of interest (unsupervised adaptation).

5 Multi-task learning

Our models perform sequence labeling based on a stacked k-layer bi-directional LSTM, a variant of LSTMs (Hochreiter and Schmidhuber, 1997) that reads the input in both regular and reversed order, allowing to take into account both left and right contexts (Graves and Schmidhuber, 2005). For our task, this enables us, for example, to distinguish between coordinated nouns and clauses. This model takes as input a sequence of words (and, here, POS tags) represented by vectors (initialized randomly or, for words, using pre-trained embedding vectors). The sequence goes through an embedding layer, and we compute the predictions of the forward and backward states for the k stacked layers. At the upper level, we compute the softmax predictions for each word based on a linear transformation. We use a logistic loss.

We also investigate joint training of multiple languages and domains for discourse segmentation. We thus try to leverage languages and domains regularities by sharing the architecture and parameters through multi-task training, where an auxiliary task is a source language (resp. domain) different from the target language (resp. domain) of interest. Specifically, we train models based on hard parameters sharing (Caruana, 1993; Collobert et al., 2011; Klerke et al., 2016; Plank et al., 2016):[4] each task is associated with a specific output layer, whereas the inner layers – the stacked LSTMs – are shared across the tasks. At training time, we randomly sample data points from one task and do forward predictions. During backpropagation, we modify the weights of the shared layers and the task-specific outer layer. The model is optimized for one target task (corresponding to the development data used). Except for the outer layer, the target task model is thus regularized by the induction of auxiliary models.

6 Corpora

Table 1 summarizes statistics about the data. For English, we use four corpora, allowing us to evaluate cross-domain performance: the RST-DT (En-DT) composed of Wall Street Journal articles; the SFU review corpus[5] (En-SFU-DT) containing product reviews; the instructional corpus (En-Instr-DT) (Subba and Di Eugenio, 2009) built

---

[4]We used a modified version of (Plank et al., 2016) fixing the random seed and using standard SGD.

[5]https://www.sfu.ca/~mtaboada

Corpus	#Doc	#EDU	#Sent	#Words
En-SFU-DT	400	28,260	16,827	328,362
En-DT	385	21,789	9,074	210,584
Pt-DT	330	12,594	4,385	136,346
Es-DT	266	3,325	1,816	57,768
En-Instr-DT	176	5,754	3,090	56,197
De-DT	174	2,979	1,805	33,591
En-Gum-DT	54	3,151	2,400	44,577
Nl-DT	80	2,345	1,692	25,095

Table 1: Number of documents, EDUs, sentences and words (according to UDPipe, see Section 7).

		Mono	Cross	UDP-S	UDP-P
languages	En-DT	89.5	**62.4**	55.6	57.5
	Pt-DT	82.2	**64.0**	49.0	62.5
	Es-DT	79.3	64.3	**64.9**	53.3
	De-DT	85.1	**76.6**	69.7	68.7
	Nl-DT	-	**82.6**	80.2	76.6
domains	En-DT (news)	89.5	**63.0**	55.6	57.5
	En-SFU-DT	85.5	**81.5**	70.2	66.1
	En-Instr-DT	87.1	**77.7**	66.5	69.5
	En-Gum-DT	-	68.1	**77.2**	61.8

Table 2: Results (F_1), comparing cross-lingual and cross-domain results with UDPipe.

on instruction manuals; and the GUM corpus[6] (En-Gum-DT) containing interviews, news, travel guides and how-tos.

For cross-lingual experiments, we use annotated corpora for Spanish (Es-DT) (da Cunha et al., 2011),[7] German (De-DT) (Stede, 2004; Stede and Neumann, 2014), Dutch (Nl-DT) (Vliet et al., 2011; Redeker et al., 2012) and, for Brazilian Portuguese, we merged four corpora (Pt-DT) (Cardoso et al., 2011; Collovini et al., 2007; Pardo and Seno, 2005; Pardo and Nunes, 2003, 2004) as done in (Maziero et al., 2015).

Three other RST corpora exist, but we were not able to obtain cross-lingual word embeddings for Basque (Iruskieta et al., 2013) and Chinese (Wu et al., 2016), and could not obtain the data for Tamil (Subalalitha and Parthasarathi, 2012).

7 Experiments

Data We use the official test sets for the En-DT (38 documents) and the Es-DT (84). For the others, we randomly choose 38 documents as test set, and either keep the rest as development set (Nl-DT) or split it into a train and a development set.

Baselines As baselines at the document level, we report the scores obtained (a) when only considering the sentence boundaries predicted using UDPipe (Straka et al., 2016) (UDP-S),[8] and (b) when EDU boundaries are added after each token PoS-tagged with "PUNCT" (UDP-P), marking either an inter- or an intra-sentential boundary.

Systems As described in Section 3, our systems are either mono-lingual or mono-domain (mono), or based on a joint training across languages or domains (cross). The "mono" systems are built for the languages and domains represented by enough data (upper part of Table 1). The "cross" models are trained using multi-task learning.

Parameters The hyper-parameters are tuned on the development set: number of iterations $i \in \{10, 20, 30\}$, Gaussian noise $\sigma \in \{0.1, 0.2\}$, and number of dimensions $d \in \{50, 500\}$. We fix the number n of stacked hidden layers to 2 and the size of the hidden layers h to 100 after experimenting on the En-DT.[9] Our final models use $\sigma = 0.2$ and $d = 500$.

Representation We use tokens and POS tags as input data.[10] The aim is to build a representation considering the current word and its context, i.e. its POS and the surrounding words/POS. We use the pre-trained UDPipe models to postag the documents for all languages. We experiment with randomly initialized and pre-trained cross-lingual word embeddings built on Europarl (Levy et al., 2017), keeping either the full 500 dimensions, or the first 50 ones.

Results Our systems are evaluated using F_1 over the boundaries (B labels), disregarding the first word of each document. Our scores are summarized in Table 2.

Our supervised, monolingual systems unsurprisingly give the best performance, with F_1 above 80%. The results are generally linked to the size of the corpora, the larger the better. Only exception is the En-SFU-DT, which, however, include more varied annotation (the authors stated that the annotations "have not been checked for reliability").

The (semi-supervised) cross-domain setting allows us to present the scores one can expect when

---

[6] https://corpling.uis.georgetown.edu/gum/
[7] We only use the test set from the annotator A.
[8] http://ufal.mff.cuni.cz/udpipe

[9] With $n \in \{1, 2, 3\}$ and $h \in \{100, 200, 400\}$).
[10] A document is a sequence alternating words and POS. The tokens are labeled with a B or an I, the POS, always labeled with an I, are inserted after each token they refer to.

only 25 documents are annotated for a new domain (i.e. the development set for the target domain), and to give the first results on the En-Gum-DT, but here, our model is actually outperformed by the sentence-based baseline (UDP-S).

The (unsupervised) cross-lingual models are generally largely better than UDPipe. These are scores that one can expect when doing cross-lingual transfer to build a discourse segmenter for a new language for which no annotated data are available. The performance is still quite high, demonstrating the coherence between the annotation schemes, and the potential of cross-lingual transfer. We acknowledge that this is a small set of relatively similar Indo-European languages, however.

Note that the sentence-based baseline has a high precision (e.g. 96.6 on Es-DT against 59.8 for the cross-lingual system), but a much lower recall, since it mainly predicts the sentence boundaries. On corpora that mostly contain sentential EDUs (e.g. Nl-DT, see Table 1), this is a good strategy. Using the punctuation (UDP-P) could be a better approximation for corpora with more varied EDUs, see the large gain for the Pt-DT and the En-Instr-DT.

Our scores are not directly comparable with sentence-level state-of-the-art systems (see Section 2). However, for En-DT, our best system correctly identifies 950 sentence boundaries out of 991, but gets only 84.5% in F_1 for intra-sentential boundaries,[11] thus lower than the state-of-the-art (91.0%). This is because we consider much less information, and because the system was not optimized for this task. Interestingly, our simple system beats HILDA (Hernault et al., 2010) (74.1% in F_1), is as good as the other neural network based system (Subba and Di Eugenio, 2007), and is close to SPADE (Soricut and Marcu, 2003) (85.2% in F_1) (Joty et al., 2015), while all of these systems use parse tree information.

Finally, looking at the errors of our system on the En-DT, we found that most of them are on the tokens "to" (30 out of 94 not predicted as 'B') and "and" (24 out of 103), as expected given the annotation guidelines (see Section 3). These words are highly ambiguous regarding discourse segmentation (e.g. in the test set, 42.3% of "and" indicates a boundary). We also found errors with coordinated verb phrases – e.g. "[when rates are rising] [**and** shift out at times]" – that should be split (Carlson et al., 2001), a distinction hard to make without syntactic trees. Finally, since we use predicted POS tags, our system learns from noisy data and makes errors due to postagging and tokenisation errors.

8 Conclusion

We proposed new discourse segmenters with good performance for many languages and domains, at the document level, within a fully predicted setting and using only language independent tools.

Acknowledgements

References

Paula C.F. Cardoso, Erick G. Maziero, Mara Luca Castro Jorge, Eloize R.M. Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago 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*. pages 88–105.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical report, University of Southern California Information Sciences Institute.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Rich Caruana. 1993. Multitask learning: a knowledge-based source of inductive bias. In *Proceedings of ICML*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research* 12:2493–2537.

Sandra Collovini, Thiago I Carbonel, Juliana Thiesen Fuchs, Jorge César Coelho, Lúcia Rino, and Renata Vieira. 2007. Summ-it: Um corpus anotado com informaçoes discursivas visandoa sumarizaçao automática. In *Proceedings of TIL*.

Iria da Cunha, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberas, and Irene Castellón.

2010. DiSeg: Un segmentador discursivo automático para el español. *Procesamiento del lenguaje natural* 45:145–152.

Iria da Cunha, Eric SanJuan, Juan-Manuel Torres-Moreno, Marina Lloberes, and Irene Castellón. 2012. DiSeg 1.0: The first system for Spanish discourse segmentation. *Expert Syst. Appl.* 39(2):1671–1678.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish Treebank. In *Proceedings of LAW*.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of ACL*.

Seeger Fisher and Brian Roark. 2007. The utility of parse-derived features for automatic discourse segmentation. In *Proceedings ACL*.

Alex Graves and Jrgen Schmidhuber. 2005. Frame-wise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks* pages 5–6.

Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse* 1:1–33.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Mikel Iruskieta, María J. Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and Oier Lopez de la Calle. 2013. The RST Basque Treebank: an online search interface to check rhetorical relations. In *Proceedings of the 4th Workshop RST and Discourse Studies*.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of ACL*.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics* 41:3.

Shafiq R. Joty, Giuseppe Carenini, Raymond T. Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of ACL*.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of NAACL*.

Omer Levy, Anders Søgaard, and Yoav Goldberg. 2017. A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of EACL*.

Jiwei Li, Rumeng Li, and Eduard H. Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of EMNLP*.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8:243–281.

Erick G. Maziero, Graeme Hirst, and Thiago A. S. Pardo. 2015. Adaptation of discourse parsing models for Portuguese language. In *Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS)*.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Cristina Bosco, Gosse Bouma, Sam Bowman, Gülşen Cebirolu Eryiit, Giuseppe G. A. Celano, Çar Çöltekin, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gokirmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Normunds Grūzītis, Bruno Guillaume, Jan Hajič, Dag Haug, Barbora Hladká, Radu Ion, Elena Irimia, Anders Johannsen, Hüner Kaşkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Simon Krek, Veronika Laippala, Lucia Lam, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Keiko Sophie Mori, Shunsuke Mori, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalnia, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Loganathan Ramasamy, Laura Rituma, Rudolf Rosa, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jing Xian Wang, Jonathan North Washington, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2016. Universal dependencies 1.3. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. http://hdl.handle.net/11234/1-1699.

Thiago A. S. Pardo and Maria das Graças Volpe Nunes. 2003. A construção de um corpus de textos científicos em Português do Brasil e sua marcação

retórica. Technical report, Universidade de São Paulo.

Thiago A. S. Pardo and Maria das Graças Volpe Nunes. 2004. Relações retóricas e seus marcadores superficiais: Análise de um corpus de textos científicos em Português do Brasil. *Relatório Técnico NILC* .

Thiago A. S. Pardo and Maria das Graças Volpe Nunes. 2008. On the development and evaluation of a Brazilian Portuguese discourse parser. *Revista de Informática Teórica e Aplicada* 15(2):43–64.

Thiago A. S. Pardo and Eloize R. M. Seno. 2005. Rhetalho: Um corpus de referłncia anotado retoricamente. In *Proceedings of Encontro de Corpora*.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of ACL*.

Gisela Redeker, Ildik Berzlnovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. Multi-layer discourse annotation of a dutch text corpus. In *Proceedings of LREC*.

Uladzimir Sidarenka, Andreas Peldszus, and Manfred Stede. 2015. Discourse segmentation of german texts. *Journal of Language Technology and Computational Linguistics* 30(1):71–98.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of NAACL*.

Caroline Sporleder and Mirella Lapata. 2005. Discourse chunking and its application to sentence compression. In *Proceedings of HLT/EMNLP*.

Manfred Stede. 2004. The potsdam commentary corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of LREC*.

Milan Straka, Jan Hajič, and Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of LREC*.

C N Subalalitha and Ranjani Parthasarathi. 2012. An approach to discourse parsing using sangati and Rhetorical Structure Theory. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012)*.

Rajen Subba and Barbara Di Eugenio. 2007. Automatic discourse segmentation using neural networks. In *Workshop on the Semantics and Pragmatics of Dialogue*.

Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of ACL-HLT*.

Nynke van der Vliet. 2010. Syntax-based discourse segmentation of Dutch text. In *15th Student Session, ESSLLI*.

Nynke Van Der Vliet, Ildikó Berzlnovich, Gosse Bouma, Markus Egg, and Gisela Redeker. 2011. Building a discourse-annotated Dutch text corpus. In *S. Dipper and H. Zinsmeister (Eds.), Beyond Semantics, Bochumer Linguistische Arbeitsberichte 3*. pages 157–171.

Yunfang Wu, Fuqiang Wan, Yifeng Xu, and Xueqiang Lü. 2016. A new ranking method for Chinese discourse tree building. *Acta Scientiarum Naturalium Universitatis Pekinensis* 52(1):65–74.

Ngo Xuan Bach, Nguyen Le Minh, and Akira Shimazu. 2012. A reranking model for discourse segmentation using subtree features. In *Proceedings of Sigdial*.