Cross-lingual RST Discourse Parsing

Chloé Braud[1], **Maximin Coavoux**[2] **and Anders Søgaard**[1]

[1]CoAStaL, DIKU, University of Copenhagen, University Park 5, 2100 Copenhagen
[2]LLF, CNRS, Univ Paris Diderot, Sorbonne Paris Cité
{braud,soegaard}@di.ku.dk
{maximin.coavoux}@etu.univ-paris-diderot.fr

arXiv:1701.02946v1 [cs.CL] 11 Jan 2017

Abstract

Discourse parsing is an integral part of understanding information flow and argumentative structure in documents. Most previous research has focused on inducing and evaluating models from the English RST Discourse Treebank. However, discourse treebanks for other languages exist, including Spanish, German, Basque, Dutch and Brazilian Portuguese. The treebanks share the same underlying linguistic theory, but differ slightly in the way documents are annotated. In this paper, we present (a) a new discourse parser which is simpler, yet competitive (significantly better on 2/3 metrics) to state of the art for English, (b) a harmonization of discourse treebanks across languages, enabling us to present (c) what to the best of our knowledge are the first experiments on cross-lingual discourse parsing.

1 Introduction

Documents can be analyzed as sequences of hierarchical discourse structures. Discourse structures describe the organization of documents in terms of discourse or rhetorical relations. For instance, the three discourse units below can be represented by the tree in Figure 1, where a relation COMPARISON holds between the segments 1 and 2, and a relation ATTRIBUTION links the segment covering the units 1 and 2, and the segment 3.[1]

1 Consumer spending in Britain rose 0.1% in the third quarter from the second quarter

2 and was up 3.8% from a year ago,

3 the Central Statistical Office estimated.

[1]"NS" and "NN" in Figure 1 describe the nuclearity of the segments, see Section 3.

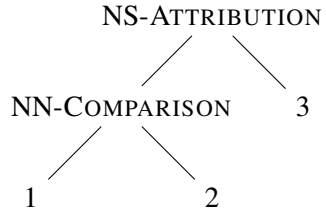

Figure 1: Tree for the structure covering the segments 1 to 3 in document 1384 in the English RST Discourse Treebank.

Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is a prominent linguistic theory of discourse structures, in which texts are analyzed as constituency trees, such as the one in Figure 1. This theory guided the annotation of the RST Discourse Treebank (RST-DT) (Carlson et al., 2001) for English, from which several text-level discourse parsers have been induced (Hernault et al., 2010; Joty et al., 2012; Feng and Hirst, 2014; Li et al., 2014; Ji and Eisenstein, 2014). Such parsers have proven to be useful for various downstream applications (Daumé III and Marcu, 2009; Burstein et al., 2003; Higgins et al., 2004; Thione et al., 2004; Sporleder and Lapata, 2005; Taboada and Mann, 2006; Louis et al., 2010; Bhatia et al., 2015).

There are discourse treebanks for other languages than English, including Spanish, German, Basque, Dutch, and Brazilian Portuguese. However, most research experimenting with these languages has focused on rule-based systems (Pardo and Nunes, 2008; Maziero et al., 2011) or has been limited to intra-sentential relations (Maziero et al., 2015).

Moreover, all discourse corpora are limited in size, since annotation is complex and time consuming. This data sparsity makes learning hard, especially considering that discourse parsing involves several complex and interacting factors, ranging from syntax and semantics, to pragmat-

ics. We thus propose to harmonize existing corpora in order to leverage information by combining datasets in different languages.

Contributions In this paper, we propose a new discourse parser that is significantly better than existing parsers for English on 2/3 standard metrics. Our parser relies on fewer features than previous work and is arguably algorithmically simpler. Moreover, we present the first end-to-end statistical discourse parsers for other languages than English (6 languages, in total). We also present the first experiments in cross-lingual discourse parsers, showing that discourse parsing is possible even when no or very little labeled data is available for the language of interest. We do so by harmonizing available discourse treebanks, enabling us to apply models across languages. We make the code and preprocessing scripts available for download at `https://bitbucket.org/chloebt/discourse`.

2 Related Work

The first text-level discourse parsers were developed for English, relying mainly on hand-crafted rules and heuristics (Marcu, 2000a; Carlson et al., 2001). Hernault et al. (2010, HILDA) greedily use SVM classifiers to make attachment and labeling decisions, building up a discourse tree. Joty et al. (2012, TSP) build a two-stage parsing system, training separate sequential models (CRF) for the intra- and the inter-sentential levels. These models jointly learn the relation and the structure, and a CKY-like algorithm is used to find the optimal tree. Feng and Hirst (2014) use CRFs only as local models for the inter- and intra-sententials levels. For Brazilian Portuguese, for example, the first system, called DiZer (Pardo and Nunes, 2008; Maziero et al., 2011), was also rule-based, but there has been some work on using classification of intra-sentential relations (Maziero et al., 2015).

Recently studies have focused on building good representations of the data. Feng and Hirst (2012) introduced linguistic features, mostly syntactic and contextual ones. Li et al. (2014) used a recursive neural network that builds a representation for each clause based on the syntactic tree, and then apply two classifiers as in Hernault et al. (2010). This leads to the best performing system for unlabeled structure (85.0 in F_1). The system presented by Ji and Eisenstein (2014, DPLP) jointly learns the representation and the task: a large margin classifier is used to learn the actions of a shift-reduce parser, optimizing at the same time the loss of the parser and a projection matrix that maps the bag-of-word representation of the discourse units into a new vector space. This system, however, only slightly outperforms the original bag-of-word representation. DPLP is the best performing discourse parser for labeled structure, 71.13 in F_1 for nuclearity and 61.63% for relation.

Our system is similar to these last approaches in learning a representation using a neural network. However, we found that good performance can already be obtained without using all the words in the discourse units, resulting in a parser that is faster and easier to adapt, as demonstrated in our multilingual experiments, see Section 7.

3 RST framework

Discourse analysis In building a discourse structure, the text is first segmented into elementary discourse units (EDU), mostly clauses. EDUs are the smallest discourse units (DUs). Discourse relations are then used to build DUs, recursively. A non-elementary DU is called a complex discourse unit (CDU). The structure of a document is the set of linked DUs. In this paper, we focus on the Rhetorical Structure Theory (RST), a theoretical framework proposed by Mann and Thompson (1988).

Nuclearity A DU is either a *nucleus* or a *satellite*, the nucleus being the most important part of the relation (i.e. of the text), while the satellite contains additional, less important information. In general, this feature depends on the relation: a relation can be either mono-nuclear (with a scheme nucleus-satellite or satellite-nucleus depending on the relative order of the spans), or multi-nuclear. Some relations can be either mono- or multi-nuclear, such as *consequence* or *evaluation* in the RST-DT.

Binary trees In the original RST framework, each relation is associated with an application scheme that defines the nuclearity of the DUs (mono- or multi-nuclear relation), and the number of DUs linked. Among the six schemes, two correspond to a link between more than two DUs, either a nucleus shared between two mono-nuclear relations (e.g. *motivation* and *enablement*) or a relation linking several nuclei (e.g. *list*). Marcu (1997) proposed to simplify the representation to

Corpus	#Doc	#Trees	#Words	#Rel	#Lab	#EDU	max/min/avg	#CDU
En-DT	385	**385**	206,300	56	110	21,789	304/2/56.6	21,404
Pt-DT	330	**329**	135,820	32	58	12,573	187/3/38.2	12,244
Es-DT[a]	266	**266**	69,787	29	43	4,019	77/2/11.5	3,671
De-DT	174	**173**	32,274	30	46	2,790	24/10/16.1	2,617
Nl-DT	80	80	27,920	31	51	2,345	47/14/29.3	2,265
Eu-DT	88	85	27,982	31	50	2,396	68/3/28.2	2,311

Table 1: Number of documents (#Doc), trees (#Trees, less than #Doc when we were unable to parse a document, see Section 4.2), words (#Words, see Section 6), relations (#Rel, originally), labels (#Lab, relation and nuclearity), EDUs (#EDU, max/min/avg number of EDUs per document), and CDUs (#CDU).

[a]The test set contains 84 documents doubly annotated, we report figures for annotator A.

binary trees, and all discourse parsers are built on a binary representation.

4 Data

We test our discourse parser on six languages, using available RST corpora harmonized as described in Section 4.2. Information about the datasets are summarized in Table 1.

4.1 RST corpora

English The RST Discourse Treebank (Carlson and Marcu, 2001), from now on En-DT, is the most widely used corpus to build discourse parsers. It contains 385 documents in English from the Wall Street Journal. The relation set contains 56 relations (ignoring nuclearity and embedding information[2]). The inter-annotator agreement scores are 88.70 for the unlabeled structure (score "Span"), 77.72 for the structure with nuclearity ("Nuclearity") and 65.75 with relations ("Relation").[3]

Brazilian Portuguese We merged all the corpora annotated for Brazilian Portuguese, as in (Maziero et al., 2015), to form the Pt-DT. The largest corpus is CST-News[4] (Cardoso et al., 2011), it is composed of 140 documents from the news domain annotated with 31 relations. Authors report agreement scores corresponding to nuclearity (0.78 in F_1) and relations (0.66).

The other corpora are: Summ-it[5] (Collovini et al., 2007) – 50 texts from science articles in a newspaper, annotated with 29 relations; Rhetalho[6] (Pardo and Seno, 2005) – 40 texts from the computer science and news domains, annotated with 23 relations; and CorpusTCC[6] (Pardo and Nunes, 2003; Pardo and Nunes, 2004) – 100 introductions of scientific texts in computer science, annotated with 31 relations.

Spanish The Spanish RST DT[7] (da Cunha et al., 2011), from now on Es-DT, contains 267 texts written by specialists on different topics (e.g. astrophysics, economy, law, linguistics) The relation set contains 29 relations. The authors report inter-annotator agreement of 86% in precision for the unlabeled structure, 82.46% for the structure with nuclearity and 76.81% with relations.

German The Postdam Commentary Corpus 2.0[8] (Stede, 2004; Stede and Neumann, 2014), from now on De-DT, contains newspaper commentaries annotated at several levels. A part of this corpus (MAZ) contains 175 documents annotated within the RST framework using 30 relations.[9]

Dutch The corpus for Dutch (Vliet et al., 2011; Redeker et al., 2012), from now on Nl-DT, contains 80 documents from expository (encyclopedias and science news website) and persuasive (fund-raising letters and commercial advertisements) genres, annotated with 31 relations. The authors report an agreement of 0.83 for discourse spans, 0.77 for nuclearity and 0.70 for relations.

[2]In this corpus, the embedded relations are annotated with a specific label (suffix "-e") that we removed.

[3]See Section 6 for a description of these metrics.

[4]http://nilc.icmc.usp.br/CSTNews/login/?next=/CSTNews/

[5]http://www.inf.pucrs.br/ontolp/downloads-ontolpplugin.php

[6]http://conteudo.icmc.usp.br/pessoas/taspardo/Projects.htm

[7]http://corpus.iingen.unam.mx/rst/index_en.html

[8]http://angcl.ling.uni-potsdam.de/resources/pcc.html

[9]We systematically ignore the first segment of each document, the title, that is not linked to the rest of the text.

Basque The Basque RST DT[10] (Iruskieta et al., 2013), from now on Eu-DT, contains 88 abstracts from three specialized domains – medicine, terminology and science –, annotated with 31 relations. The inter-annotator agreement is 81.67% for the identification of the CDU (Iruskieta et al., 2015), and 61.47% for the identification of the relations.

Other corpora To the best of our knowledge, the only two non English corpora not included are the one annotated for Tamil (Subalalitha and Parthasarathi, 2012) that we were unable to find, and the (intra-sentential) one developed for Chinese (Wu et al., 2016), for which we were unable to produce RST trees since annotation does not contain nuclearity indications.

For English, there are corpora annotated for other domains than the one covered by the En-DT. We however leave out-of-domain evaluation for future work: it requires to decide how to use a corpus annotated only at the sentence level (SFU review corpus)[11], or a corpus annotated with genre specific relations (Subba and Di Eugenio, 2009).

4.2 Harmonization of the datasets

Recent discourse parsers built on the En-DT are based on pre-processed data: the corpus contains only binary trees, with the large label set mapped to 18 coarse-grained classes. In this section, we describe this pre-processing step for all corpora used. Discourse corpora have been released under three different file formats: `dis` (En-DT), `lisp` (Rhetalho and CorpusTCC) and `rs3` (all remaining corpora). The first two ones are bracketed format, the third one is an XML encoding. In all cases, the trees encoded do not look like the one in Figure 1: the relations are annotated on the daughter nodes, on the satellite for mono-nuclear relations, or on all the nuclei for multi-nuclear relations. Moreover, in the `rs3` format, the nuclearity of the segments is not directly annotated, it has to be retrieved using the type of the relation (indicated at the beginning of each file) and the previous principle. Our pre-processing step leads to corpora with bracketed files representing directly the RST trees (as in Figure 1) with stand-off annotation of the text of the EDUs.

Note that, even if harmonized, the corpora are not parallel, making it hard to use them to study language variations for the discourse level. Some preliminary work exists on this question (Iruskieta et al., 2015).

Pre-processing Some documents (format `rs3`) contain several roots or empty segments. We were generally able to remove useless units, that is units that are not linked to other ones within the tree, except for one document in the CST corpus (two roots, both linked to other units).

Another issue concerns unordered EDUs: the structure annotated contains nodes spanning non adjacent EDUs. In general, we were able to correct these cases, but we failed to automatically produce trees spanning only adjacent EDUs for three documents in the Eu-DT, and one document in the De-DT.

Binarization All the corpora contain non-binary trees that we map to binary ones. In the En-DT, common cases of non-binarity are nodes whose daughters all hold the same multi-nuclear relation – indicating that this relation spans multiple DUs, e.g. *list*.[12] In rare cases, the children are two satellites and a nucleus – indicating that the nucleus is shared by the satellites. These configurations are the ones described in (Marcu, 1997) (see Section 3), and choosing right or left-branching leads to a similar interpretation. For the En-DT, right-branching is the chosen strategy since (Soricut and Marcu, 2003).

We found more diverse cases in the other corpora, and, for some of them, right-branching is impossible. It is the case when the daughters are one nucleus (annotated with "Span", only indicating that this node spans several EDUs) and more than two satellites holding different relations – i.e. the nucleus is shared by all the relations. More precisely, the issue arises when the last two children are satellites. Using right-branching, we end with a node with two satellites as daughters, and thus a ill-formed tree. In order to keep as often as possible the "right-branching by default" strategy, we first do a right-branching and then a left-branching: beginning with four children – S_1-R_i, N_2-Span, S_3-R_j and S_4-R_k, indicating the relations $R_i(S_1, N_2)$, $R_j(N_2, S_3)$ and $R_k(N_2, S_4)$[13] –, we end up with the tree in Figure 2. Finally, we used a right-branching in all cases, except when the two last children are satellites.

[10] `http://ixa2.si.ehu.es/diskurtsoa/en/`

[11] `https://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html`

[12] Recall that in the original format, the relation is not annotated on the parent node but on the children.

[13] S being a satellite, N a nucleus and R a relation.

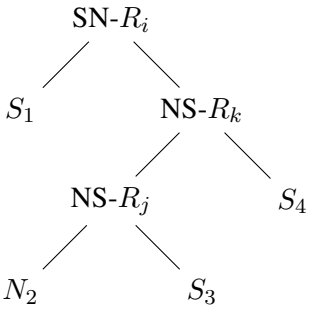

Figure 2: Binary tree for a node X with 4 children: S_1-R_i, N_2-Span, S_3-R_j and S_4-R_k.

Label set harmonization We map all the relations used in the corpora to the 18 coarse grained classes (Carlson and Marcu, 2001) used to build the most recent discourse parsers on the En-DT.[14]

The mapping for the En-DT is given in (Carlson and Marcu, 2001). For all the other corpora, we first map all the relations that exist in this mapping (i.e. used in the En-DT annotation scheme) to their corresponding classes. We end with 18 problematic relations, that is relations that were not used when annotating the En-DT.

Among them, 10 can be mapped easily, because they directly correspond to a class – *explanation* is mapped to the class EXPLANATION, *elaboration* to ELABORATION, *joint* to JOINT –, because they were just renamed – *reformulation* is mapped to the class RESTATEMENT and *solutionhood* (same as *problem-solution*) to TOPIC-COMMENT –, or because they correspond to a more-fine grained formulation of existing relations – *entity-elaboration* is mapped to ELABORATION and the 4 *volitional/non-volitional cause* and *result* are mapped to the class CAUSE, corresponding to the relations *cause* and *result* in the En-DT.

For the remaining relations, we looked at the definition of the relations[15] to decide on a mapping. Note that this label mapping is made quite easy by the fact that all the corpora were annotated following the same underlying theory – they thus use relations defined using similar criteria –, and that we are using a coarse-grained classification – we thus do not need to decide whether a relation is equivalent to another one, but rather whether it fits the properties of the other relations within a specific class. Label mappings for corpora annotated following different frameworks are still discussed (Roze, 2013; Benamara and Taboada, 2015).

We decided on the following mapping, considering the properties of the relations and the classes: *parenthetical* – used to give "additional details" – is mapped to ELABORATION, *conjunction* – similar to a *list* with only two elements – to JOINT, *justify* – similar to *Explanation-argumentative* – and *motivation* – quite similar to *reason* and grouped with *evidence* in (Benamara and Taboada, 2015) – to EXPLANATION, *preparation* – presenting preliminary information, increasing the readiness to read the nucleus – to BACKGROUND, and *unconditional* and *unless* – linked to *condition* – to CONDITION.

Finally, note that this mapping does not lead to having the same relation set for all the corpora, and that the relation distribution could vary among the datasets.

5 Discourse Parser

Our discourse parser builds discourse structures from segmented texts, we did not implement discourse segmenters for each language. Discourse segmenters only exist for English (Hernault et al., 2010) (95, 0% in F_1), Brazilian Portuguese (Pardo and Nunes, 2008) (56.8%) and Spanish (da Cunha et al., 2010; da Cunha et al., 2012) (80%). Discourse segmenters can be built quite easily relying only on manual rules as it is the case for the Spanish and Portuguese ones, especially considering that segmentation has generally been made coarser in the corpora built after the En-DT (Vliet et al., 2011). While improving this first step is crucial, we focus on the harder step of tree building.

5.1 Description of the Parser

We used the syntactic parser described in Coavoux and Crabbé (2016), in the static oracle setting. We chose this parser because it can take pre-trained embeddings as input and, more importantly, because it was designed for morphologically rich languages and thus takes as input not only tokens and POS tags, but any token attribute that is then mapped to a real-valued vector, which allows the use of complex features.

The parser is a transition-based constituent parser that uses a lexicalized shift-reduce transition system (Sagae and Lavie, 2005). The transition system is based on two data structures – a *stack* (S) stores partial trees and a *queue* (B) con-

[14] The full mapping is provided in Appendix A.

[15] http://www.sfu.ca/rst/01intro/definitions.html

tains the unparsed DUs. A parsing *configuration* is a couple $\langle S, B \rangle$. In the initial configuration, S is empty and B contains the whole document. The parser iteratively applies actions to the current configuration, in order to derive new configurations until it reaches a final state, i.e. a parsing configuration where B is empty and S contains a single element (the root of the tree).

The actions are defined as follows:

- SHIFT pops an EDU from B and pushes it onto S.
- REDUCE-R-X and REDUCE-L-X pop two DUs from S, push a new CDU with the label X on S and assign its nucleus (Left or Right).

Scoring System As in Chen and Manning (2014), at each parsing step, the parser scores actions with a feed-forward neural network. The input of the network is a sequence of typed symbols extracted from the top elements of S and B. The symbols are typically discourse relations or attributes of their nucleus EDU (e.g. first word of EDU, see Section 5.3).

The first layer of the network projects these symbols onto an embedding space (each type of symbol has its own embedding matrix). The following two layers are non-linear layers with a ReLU activation. The output of the network is a probability distribution over possible actions computed by a softmax layer.

To generate a set of training examples $\{a^{(i)}, c^{(i)}\}_{i=1}^{N}$, we used the static oracle to extract the gold sequence of actions and configurations for each tree in the corpus. The objective function of the parser is the negative log-likelihood of gold actions given corresponding configurations:

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{i=1}^{N} \log P(a^{(i)}|c^{(i)}; \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is the set of all parameters, including embedding matrices.

We optimized this objective with the averaged stochastic gradient descent algorithm (Polyak and Juditsky, 1992). At inference time, we used beam-search to find the best-scoring tree.

5.2 Cross-lingual Discourse Parsing

Our first experiments are strictly monolingual, and they are intended to give state-of-the-art performance in a fully supervised setting. We consider that we need at least 100 documents to build a monolingual model, since we already keep around 65 documents for test and development.

We then evaluate multi-source transfer methods, considering one language as the target and the others as sources. More precisely, we will evaluate two settings: (1) training and optimizing only on the available source data; (2) training on all available data, including target ones if any, and optimizing on the development set of the target language. Setting (1) provides performance when no data are available at all in the target language, while (2) aims at evaluating if one can expect improvements by simply combining all the available data.

When combining the corpora, we cannot ignore lexical information as it has been done for syntactic parsing with delexicalized models (McDonald et al., 2011). Discourse parsing is a semantic task, at least when it comes to predict a rhetorical relation between two spans of text, and information from words have proven to be crucial (Rutherford and Xue, 2014; Braud and Denis, 2015). We thus include word features using bilingual dictionaries – i.e. translating the words used as features into a single language (English) –, or through cross-lingual word embeddings as proposed in (Guo et al., 2015) for dependency parsing. More precisely, we used the cross-lingual word representations presented in (Levy et al., 2017) that allow multi-source learning and have proven useful for POS tagging but also more semantic-oriented tasks, such as dependency parsing and document classification.

5.3 Features

As in previous studies, we used features representing the two EDUs on the top of the stack and the EDU on the queue. If the stack contains CDUs, we use the nuclearity principle to choose the head EDU, converting multi-nuclear relations into nucleus-satellite ones as done since (Sagae, 2009). However, we found that using these information also for the left and right children of the two CDUs on the top of the stack, and adding as a feature the representation built for these two CDUs lead to important improvements.

Lexical features We use the first three words and the last word along with their POS, features that have proven useful for discourse (Pitler et al., 2009), and the words in the *head set* (Sagae, 2009)

– i.e. words whose head in the dependency graph is not in the EDU –, here limited to the first three.[16] This head set contains the head of the sentence (in general, the main event), or words linked to the main clause when the segment does not contain the head (especially, discourse connectives that are subordinating or coordinating conjunctions could be found there). The words are the boundaries could also contain discourse connectives, adverbs or temporal expressions that could be relevant for discourse structure. Note however that these feature have been built for English, and they could be less useful for other languages. We leave the question of investigating their utility linked to word order differences for future work.

Note that we do not use all the words in the EDUs as features, contrary to (Li et al., 2014; Ji and Eisenstein, 2014). Our only word features are the words in the head set and at the boundaries, thus 7 words per EDU. When using word embeddings, we concatenate the vectors for each word, each of d dimensions, keeping the same order to build a vector of $7d$ dimensions (e.g., the first word of the EDU corresponds to the first d dimensions, the second has values between d and $2d$).

Position and length Other features are used to represent the position of the EDU in the document and its length in tokens. We use thresholds to distinguish between very long (length $l > 25$ tokens), long ($l > 15$), short ($l > 5$) and very short ($l \leq 5$) EDUs. We also distinguish between the "first" and the "last" EDU in the document, and use also a threshold on the ratio $s =$(position of the EDU divided by the total number of EDUs) to separate EDUs at the beginning ($s < 0.25$), in the first middle ($0.25 \leq s < 0.5$), in the second middle ($0.5 \leq s < 0.75$) or in the end ($s >= 0.75$).

Position of the head We add a boolean feature indicating if the head of the sentence is in the current EDU or outside.

Number/date/percent/money We also use 4 indicators of the presence of a date, a number, an amount of money and a percentage, features that have proven to be useful for discourse (Pitler et al., 2009). We build these features using simple regular expressions.

Corpus	Size dict.	# words	# unk. words
Pt-DT	18,049	13,417	6,929
Es-DT	22,815	6,961	3,231
De-DT	31,900	5,856	1,762
Nl-DT	19,012	3,316	1,428
Eu-DT	1,092	6,553	5,446

Table 2: Dictionary coverage for each dataset on the train set when available, on the dev set else.

6 Experiment settings

Data For the En-DT, we follow previous works in using the official test set of 38 documents. For the Es-DT, we report results on the test set A.[17] For all the other corpora, we randomly choose 38 documents to make a test set, and either use the remaining documents as development set (Nl-DT and Eu-DT), or split them into a development set of 25 documents, the remaining being used as training set (En-DT, Es-DT, Pt-DT and De-DT).

All the results given are based on a gold segmentation of the documents.

Each dataset is parsed using UDPipe,[18] thus tokenizing, splitting into sentences and annotating each document based on the Universal Dependency scheme (Nivre et al., 2016).

The word features for the non-English datasets are translated using available bilingual Wiktionaries[19] without disambiguation, the coverage of each dictionary is given in Table 2. We also look for a translation of the lemma (and of the stems for the languages for which a stemmer[20] was available) as a backup strategy. When no translation is found, we keep the original token.

The word embeddings used were built on the EuroParl corpus (Levy et al., 2017). We keep only the 50 first dimensions of the vectors representing the words, our preliminary experiments suggesting no significant differences against keeping the whole 200 dimensions. Unknown words are represented by the average vector of all word vectors. For Basque, we had no access to these embeddings, we thus only report results using bilingual Wiktionaries.

[16] Having more than three tokens in the head set is rare.

[17] We found similar performance on the other test set.

[18] http://ufal.mff.cuni.cz/udpipe

[19] https://en.wiktionary.org/wiki/User:Matthias_Buchmeier

[20] https://pypi.python.org/pypi/snowballstemmer

Parameter tuning In our experiments we optimized on the development set the following parameters: the learning rate $\in \{0.01, 0.02, 0.03\}$, the learning rate decay constant $\in \{10^{-5}, 10^{-6}, 10^{-7}, 0\}$, the number of iterations $\in [1 - 20]$, and the size of the beam $\in \{1, 2, 4, 8, 16, 32\}$. We fixed the number N of hidden layers to 2 and the size of the hidden layers H to 128 after experimenting on the En-DT (with $N \in \{1, 2, 3\}$ and $H \in \{64, 128\}$).

We fixed the size of the vectors for each feature to 50 for word features,[21] 16 for POS, 6 for position, 4 for length, and 2 for other features.

Metrics Following (Marcu, 2000b) and most subsequent work, output trees are evaluated against gold trees in terms of how similar they bracket the EDUs (Span), how often they agree about nuclei when predicting a true bracket (Nuclearity), and in terms of the relation label, i.e., the overlap between the shared brackets between predicted and gold trees (Relation).[22] These scores are analogous to labeled and unlabeled syntactic parser evaluation metrics.

Baseline Since we do not have state-of-the-art results for most of the languages, we provide results for a simple most frequent baseline (System MFS) that labels all nodes with the most frequent relation in the training or development set – that is NN-JOINT for De-DT and Es-DT, and NS-ELABORATION for the others –, and build the structure by right-branching.

7 Results

Monolingual experiments Monolingual experiments are aimed at evaluating performance for languages having a large annotated corpus (at least 100 documents). Our results are summarized in Table 3. Our parser is competitive with state-of-the-art systems for English (first line in Table 3), with even better performance for unlabeled structure (85.04%) and structure labeled with nuclearity (72.29%). These results show that using all the words in the units (Ji and Eisenstein, 2014; Li et al., 2014), is not as useful as using more contextual information, that is taking more DUs into account (left and right children of the CDUs in the stack). However, the slight drop for Relation shows that we probably miss some lexical information, or that we need to choose a more effective combination scheme than concatenation. We plan to use bi-LSTM encoders (Hochreiter and Schmidhuber, 1997) to construct fixed-length representations of EDUs.

For the other languages, performance are still high for unlabeled structure, but far lower for labeled structure except for Spanish. For this language, the quite high performance obtained were unexpected, since the corpus is far much smaller than the Portuguese one. One possible explanation is that the Portuguese corpus is in fact a mix of different corpora, with varied domains, and possibly changes in annotation choices. On the other hand, the low results for German show the sparsity issue since it is the language for which we have the fewest annotations ("#CDU", see Table 1).

Cross-lingual experiments When only relying on data from different languages ("Cross" in Table 3), we observe a large drop in performance compared to monolingual systems. The source-only discourse parsers still have fairly high performance for unlabeled structure (around 70% or higher), the scores being especially low for relation identification. This could indicate that our representation does not generalize well. But it also comes from differences among the corpora. For example, only the En-DT and the Pt-DT use the relation ATTRIBUTION. This leads to a large drop in performance associated with this relation, when one of these corpora is not in the training data, especially for the source-only system for the En-DT (from 93% in F_1 to 30%). On the other hand, on the En-DT, we observe improvement for other relations either largely represented in all the corpora (e.g. JOINT +3%), or under-represented in the En-DT (e.g. CONDITION +3%).

When combining corpora for source and target languages ("+ dev." in Table 3), we obtain our best performing system for English, with all scores improved compared to our best monolingual system (+0.8 for Nuclearity and +1.3 for Relation). Otherwise scores are similar to the monolingual case.

Finally, for languages without training set (Nl-DT and Eu-DT), this strategy allows us to build parsers outperforming our simple baseline (MFS) by around 11−13% for Span, 8−15% for Nuclearity and 6−11% for Relation. Having at least some annotated data to make a development set allows improvements against only using corpora in other

[21] When using embeddings, the final vector is of size 350.

[22] We use the evaluation script provided at `https://github.com/jiyfeng/DPLP`.

System	En-DT			Pt-DT			Es-DT			De-DT			Nl-DT			Eu-DT		
	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel	Sp	Nuc	Rel
MFS	58.2	33.4	22.1	57.3	33.9	23.23	82.0	51.5	17.7	61.3	37.8	13.2	57.9	35.5	22.0	63.2	34.9	18.8
Li et al.[a]	85.0	70.8	58.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DPLP[a]	82.1	71.1	**61.6**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mono	85.0	72.3	60.1	**82.0**	**65.1**	**49.9**	**89.7**	**72.7**	**54.4**	**80.2**	**53.9**	**35.0**	-	-	-	-	-	-
+ emb.	83.5	68.5	55.9	81.3	62.9	48.8	89.3	72.4	51.4	77.7	51.6	31.1	-	-	-	-	-	-
Cross	76.3	50.5	31.3	76.5	54.6	35.5	78.1	45.4	27.0	76.0	46.0	26.1	69.5	42.1	25.3	**78.6**	**53.0**	26.4
+ dev.	**85.1**	**73.1**	61.4	81.9	65.1	49.8	88.8	68.0	50.4	79.6	53.6	34.1	**69.2**	**43.4**	**28.3**	76.7	50.5	**29.5**
Human[b]	88.7	77.7	65.8	-	78	66	86	82.5	76.8	-	-	-	83	77	70	81.7	-	61.5

Table 3: Performance of our monolingual and cross-lingual systems for Span (Sp), Nuclearity (Nuc) and Relation (Rel). "MFS" corresponds to the baseline system described in Section 6; "+ emb." is the monolingual system using word embeddings; "+dev." means that the system is optimized on the development set of the target language (*vs* the union of the source development sets). For cross-lingual systems, we only report our best results using either word embeddings or bilingual dictionaries.

[a] Scores reported from (Li et al., 2014), and DPLP (Ji and Eisenstein, 2014).

[b] For Brazilian Portuguese, inter-annotator agreement scores are only available for the CST-news corpus ; For Spanish, only precision scores are reported ; For Basque, the scores reported are different (Iruskieta et al., 2015).

languages (around $+3\%$ for the Nl-DT and the Eu-DT for Relation). On the other hand, we probably overfit our development data for the Eu-DT, since better results were obtained for unlabeled structure ($+2\%$) and structure with nuclearity ($+2.5\%$) using only data in other languages.

Word embeddings Using word embeddings ("+emb" in Table 3) for monolingual systems often leads to an important drop in performance, especially for Relation (from -1.1 to -4.2%). This demonstrates that these embeddings do not provide the large range of information needed for relation identification, a task inherently semantic. We believe however that the results are not too low to prevent for interesting applications. It is noteworthy that the English parser with embeddings is still better than the systems proposed in (Hernault et al., 2010; Joty et al., 2013).

For cross-lingual experiments, the bilingual dictionaries perform generally better than embeddings (except for Pt-DT and De-DT for source-only systems), demonstrating again that we need representations more tailored to the task to leverage all relevant lexical information.

8 Conclusion

We introduced a new discourse parser that obtains state-of-the-art performance for English. We harmonized discourse treebanks for several languages, enabling us to present results for five other languages for which available corpora are smaller, including the first cross-lingual discourse parsing results in the literature.

Acknowledgements

We thank the three anonymous reviewers for their comments. Chloé Braud and Anders Søgaard were funded by the ERC Starting Grant LOWLANDS No. 313695.

References

[Benamara and Taboada2015] Farah Benamara and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of Starsem*.

[Bhatia et al.2015] Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of EMNLP*.

[Braud and Denis2015] Chloé Braud and Pascal Denis. 2015. Comparing word representations for implicit discourse relation classification. In *Proceedings of EMNLP*.

[Burstein et al.2003] Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the write stuff: automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Advances in Natural Language Processing*, 18.

[Cardoso et al.2011] 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.

[Carlson and Marcu2001] Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. Technical report, University of Southern California Information Sciences Institute.

[Carlson et al.2001] 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*.

[Chen and Manning2014] Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Empirical Methods in Natural Language Processing (EMNLP)*.

[Coavoux and Crabbé2016] Maximin Coavoux and Benoit Crabbé. 2016. Neural greedy constituent parsing with dynamic oracles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 172–182, Berlin, Germany, August. Association for Computational Linguistics.

[Collovini et al.2007] 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. *Proceedings of TIL*.

[da Cunha et al.2010] 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.

[da Cunha et al.2011] Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish Treebank. In *Proceedings of the Fifth Linguistic Annotation Workshop, LAW*.

[da Cunha et al.2012] 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.

[Daumé III and Marcu2009] Hal Daumé III and Daniel Marcu. 2009. A noisy-channel model for document compression. In *Proceedings of ACL*.

[Feng and Hirst2012] Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of ACL*.

[Feng and Hirst2014] Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of ACL*.

[Guo et al.2015] Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of ACL-IJCNLP*.

[Hernault et al.2010] 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.

[Higgins et al.2004] Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of HLT-NAACL*.

[Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[Iruskieta et al.2013] 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*.

[Iruskieta et al.2015] Mikel Iruskieta, Iria da Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. In *Proceedings of LREC*.

[Ji and Eisenstein2014] Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of ACL*.

[Joty et al.2012] Shafiq R. Joty, Giuseppe Carenini, and Raymond T. Ng. 2012. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of EMNLP*.

[Joty et al.2013] 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*.

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

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

[Louis et al.2010] Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of SIGDIAL*.

[Mann and Thompson1988] William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8:243–281.

[Marcu1997] Daniel Marcu. 1997. From discourse structures to text summaries. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 82–88.

[Marcu2000a] Daniel Marcu. 2000a. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*.

[Marcu2000b] Daniel Marcu. 2000b. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA, USA.

[Maziero et al.2011] Erick G. Maziero, Thiago A. S. Pardo, Iria da Cunha, Juan-Manuel Torres-Moreno, and Eric SanJuan. 2011. DiZer 2.0-an adaptable on-line discourse parser. In *Proceedings of 3rd RST Brazilian Meeting*, pages 1–17.

[Maziero et al.2015] 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)*.

[McDonald et al.2011] Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of EMNLP*.

[Nivre et al.2016] 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.

[Pardo and Nunes2003] 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, Technical Report.

[Pardo and Nunes2004] 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*.

[Pardo and Nunes2008] 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.

[Pardo and Seno2005] Thiago A. S. Pardo and Eloize R. M. Seno. 2005. Rhetalho: Um corpus de referłncia anotado retoricamente. In *Proceedings of Encontro de Corpora*.

[Pitler et al.2009] Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of ACL-IJCNLP*.

[Polyak and Juditsky1992] Boris T. Polyak and Anatoli B. Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, July.

[Redeker et al.2012] 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*.

[Roze2013] Charlotte Roze. 2013. *Vers une algèbre des relations de discours*. Ph.D. thesis, Université Paris-Diderot.

[Rutherford and Xue2014] Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of EACL*.

[Sagae and Lavie2005] Kenji Sagae and Alon Lavie. 2005. A classifier-based parser with linear run-time complexity. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 125–132. Association for Computational Linguistics.

[Sagae2009] Kenji Sagae. 2009. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of IWPT 2009*.

[Soricut and Marcu2003] Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of NAACL*.

[Sporleder and Lapata2005] Caroline Sporleder and Mirella Lapata. 2005. Discourse chunking and its application to sentence compression. In *Proceedings of HLT/EMNLP*.

[Stede and Neumann2014] Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of LREC*.

[Stede2004] Manfred Stede. 2004. The potsdam commentary corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*.

[Subalalitha and Parthasarathi2012] 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)*.

[Subba and Di Eugenio2009] Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of ACL-HLT*.

[Taboada and Mann2006] Maite Taboada and William C. Mann. 2006. Applications of rhetorical structure theory. *Discourse Studies*, 8:567–588.

[Thione et al.2004] Gian Lorenzo Thione, Martin Van den Berg, Livia Polanyi, and Chris Culy. 2004. Hybrid text summarization: Combining external relevance measures with structural analysis. In *Proceedings of the ACL Workshop Text Summarization Branches Out*.

[Vliet et al.2011] 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.

[Wu et al.2016] 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.

A Mapping of the relations

Classe	Relations
ATTRIBUTION	*attribution*, *attribution-negative*
BACKGROUND	*background*, *circumstance*, *circunstancia*, *fondo*, *preparación*, *preparation*, *prestatzea*, *testuingurua*, *zirkunstantzia*
CAUSE	*causa*, *cause*, *cause-result*, *consequence*, *kausa*, *non-volitional-cause*, *non-volitional-result*, *ondorioa*, *result*, *resultado*, *volitional-cause*, *volitional-result*
COMPARISON	*analogy*, *comparison*, *preference*, *proportion*
CONDITION	*alderantzizko-baldintza*, *alternativa*, *aukera*, *baldintza*, *condición*, *condición-inversa*, *condition*, *contingency*, *ez-baldintzatzailea*, *hypothetical*, *otherwise*, *unconditional*, *unless*
CONTRAST	*antitesia*, *antithesis*, *antítesis*, *concesión*, *concession*, *contrast*, *contraste*, *kontrastea*, *kontzesioa*
ELABORATION	*definition*, *e-elaboration*, *elaboración*, *elaboration*, *elaboration-additional*, *elaboration-general-specific*, *elaboration-object-attribute*, *elaboration-part-whole*, *elaboration-process-step*, *elaboration-set-member*, *elaborazioa*, *example*, *parenthetical*
ENABLEMENT	*ahalbideratzea*, *capacitación*, *enablement*, *helburua*, *propósito*, *purpose*
EVALUATION	*comment*, *conclusion*, *ebaluazioa*, *evaluación*, *evaluation*, *interpretación*, *interpretation*, *interpretazioa*
EXPLANATION	*ebidentzia*, *evidence*, *evidencia*, *explanation*, *explanation-argumentative*, *justificación*, *justifikazioa*, *justify*, *motibazioa*, *motivación*, *motivation*, *reason*
JOINT	*bateratzea*, *conjunción*, *conjunction*, *disjunction*, *disjuntzioa*, *disyunción*, *joint*, *konjuntzioa*, *list*, *lista*, *unión*
MANNER-MEANS	*manner*, *means*, *medio*, *metodoa*
SAME-UNIT	*same-unit*
SUMMARY	*birformulazioa*, *definitu-gabeko-erlazioa*, *laburpena*, *reformulación*, *restatement*, *resumen*, *summary*
TEMPORAL	*inverted-sequence*, *secuencia*, *sekuentzia*, *sequence*, *temporal-after*, *temporal-before*, *temporal-same-time*
TEXTUAL-ORGANIZATION	*textual-organization*
TOPIC-CHANGE	*topic-drift*, *topic-shift*
TOPIC-COMMENT	*arazo-soluzioa*, *comment-topic*, *problem-solution*, *question-answer*, *rhetorical-question*, *solución*, *solutionhood*, *statement-response*, *topic-comment*

Table 4: Mapping of all the relations found in the datasets: for each class, we give the set of relation names as they appear in the corpora (removing only the possible suffixes "-e", "-s", "-mn"). We ignore the simplest differences in names (e.g. *textual-organization* and *textualorganization*).