

Empirical validations of multilingual annotation schemes for discourse relations

Sandrine Zufferey
U.C. Louvain
Institut langage et
communication
Place B. Pascal, 1
B-1348 Louvain

sandrine.zufferey
@uclouvain.be

Liesbeth Degand
U.C. Louvain
Institut langage et
communication
Place B. Pascal, 1
B-1348 Louvain

liesbeth.degand
@uclouvain.be

Andrei Popescu-Belis
Idiap Research Institute
Rue Marconi 19
CH-1920 Martigny

andrei.popescu-
belis@idiap.ch

Ted Sanders
Universiteit Utrecht
Utrecht Institute of
Linguistics
Trans 10
NL-3512 JK Utrecht

T.J.M.Sanders@uu.nl

Abstract

This paper discusses the empirical validation of annotation schemes proposed for discourse relations, when signaled explicitly by discourse connectives, through their application to texts in several languages. Considering a monolingual annotation scheme as a starting point, the paper explains the reasons for either specifying or generalizing some labels, illustrating them with a review of experiments in translation spotting of connectives. Then, an experiment with the PDTB scheme applied to five languages (EN, FR, DE, NL, and IT) shows how specification and generalization are put to work in order to build a scheme which has an improved empirical validity for several languages.

1 Introduction

Several corpora with annotated discourse relations have become available in the past years, inspired by the first lexicalized discourse structure annotation performed for the Penn Discourse Treebank (PDTB, Prasad, Dinesh, Lee et al., 2008), which has become a landmark in the field – see Webber and Joshi (2012) for a review. These annotation efforts have reused

and sometimes redefined the annotation instructions and the classification of discourse relations proposed by the PDTB. This taxonomy holds for discourse relations that can be lexicalized through the use of discourse connectives, but also for implicit relations that are not lexicalized.

In this paper, we focus on lexicalized discourse relations, made explicit by discourse connectives, in a parallel corpus with translations from English into four other languages. Through a series of experiments with the PDTB taxonomy of discourse relations, we show how this taxonomy should be adapted to suit the needs of several languages and to make the annotation process more accurate (Sections 4 to 6). However, we initially reflect from a more general perspective on the benefits of multilingual annotation for designing a standardized taxonomy of discourse relations applicable across languages. After stating the problem theoretically (Section 2), we review monolingual and multilingual annotations, including translation spotting of discourse connectives in parallel corpora (Section 3).

2 Impact of multilingual annotations on taxonomies of discourse relations

The attempt to define a universally acceptable list of discourse relations (Bunt, Prasad and Joshi, 2012) raises several theoretical questions about the principles governing such a list. In our view, some of the most important ones are:

- What counts as a discourse relation and what theory should be used to list possible relations?
- Are discourse relations truly language-independent, i.e. can all of them be encountered in texts from any language?
- Are all discourse relations equally achievable by implicit and explicit means? In particular, are there, in a given language, connectives to express all relation types?
- What is the relation between a language-independent taxonomy of discourse relations and the range of discourse connectives available in a given language? How can such a taxonomy be used to map discourse connectives from one language to another?
- Do all discourse relations that can be expressed by a given connective count as possible meanings of that connective?
- Given that one connective is almost never fully substitutable with another one, are there more meanings than connectives? And what accounts for the diversity of connectives in European languages?

These questions are, of course, far beyond the scope of this paper. In this section, we will first state two principles that govern the relations between a taxonomy of discourse relations and the vocabularies of discourse connectives in several European languages. We will also briefly discuss the relation between semantic meaning and meaning in context for discourse connectives.

2.1 Specification vs. generalization in a taxonomy of discourse senses

Let us consider first an existing taxonomy such as the PDTB, used for the annotation of a large English corpus, and let us suppose a translation of the annotated corpus is available in French. Then, when examining all occurrences of an English discourse connective C_i annotated with a sense R_n from the taxonomy, it might happen that several different translations of C_i are observed (with significant frequencies), and that these different translations correspond to a previously uncategorized distinction of the discourse relation R_n . Hence, in this case, R_n

must be subdivided into two more specific relations, say R_{n1} and R_{n2} . We call this the *specification process* (or *refinement*) of the taxonomy.

Consider now a different case: after application to annotation over large corpora in several languages, it is found that two senses of a taxonomy, say R_{p1} and R_{p2} exhibit low inter-annotator agreement, and are often dispreferred in favor of their supersense (in the taxonomy), say R_p . In this case, it makes sense to prune the two senses from the taxonomy and keep only their supersense. Of course, this does not rule out the possibility that when a new language is annotated, the supersense must be again specified. However, until such additional evidence is found, a more compact taxonomy ensures higher inter-coder agreement. We call *abstraction* (or *generalization*) the process described above.

The main stance of this paper is that, in order to obtain a normalized scheme, one can: (1) start with a theoretically-grounded taxonomy (e.g. the PDTB, or an RST-based one), and (2) submit it to empirical testing, which means using specification and generalization to make it evolve into a truly universal, empirically-grounded multi-lingual taxonomy.

2.2 Semantic vs. contextual meanings of discourse connectives

A difficulty for the annotation of the rhetorical relations conveyed by connectives is that connectives can be used to convey a different relation than the one(s) that they semantically encode. The best-known case of this type of semantic under-determination is the connective *and*, which often conveys in context a more specific relation than its semantic meaning of addition, notably a temporal or a causal meaning (e.g. Spooren, 1997; Carston, 2002). These relations are then called its pragmatic meanings. Most analyses treat these pragmatic meanings as inferable in context but not as part of the semantic meaning of *and*. This phenomenon is also observed with other connectives; for example, temporal connectives may at times convey a causal or a contrastive relation as their pragmatic meaning, without having these relations as part of their semantic core meaning. This phenomenon is distinct from the semantic

ambiguity of connectives (such as *since*) that can alternatively convey distinct semantic meanings (for *since*, temporal or causal).

Therefore, an important question is to define what level of meaning (semantic or pragmatic) has to be annotated. Obviously, the pragmatic relation conveyed in context is more helpful for understanding the contribution of a connective in a given utterance than its core semantic meaning. However, relations that differ in context from the semantic meaning of a connective give rise to an important number of disagreements between annotators, probably because in such cases the interpretation rests on inference, a process that varies across speakers (cf. Spooen and Degand 2010).

In our view, a way to deal with the under-determinacy question is to make annotators aware of this phenomenon and encourage the annotation of the meaning perceived in context, even when it departs from the connective's core semantic meaning. However, the latter meaning must be taken into account if the annotation is used to establish the range of possible semantic meanings of discourse connectives, and in particular if frequency information is desired. This is especially the case for lexicographic analyses which look for statistics regarding semantic meanings only.

3 Previous work and results

Evidence for the applicability of the PDTB to several languages comes from recent experiments with monolingual annotations. The PDTB has indeed set the example for a number of other lexicalized, monolingual taxonomies of discourse relations (reviewed by Webber and Joshi, 2012), namely in Czech (Zikánová et al., 2010), Arabic (Al-Saif and Markert, 2010), Chinese (Huang and Chen, 2011; Zhou and Xue, 2012), Hindi (Kolachina et al., 2012) and Turkish (Zeyrek et al., 2010). An annotation project aiming at a French Discourse Treebank is also in progress (Danlos et al., 2012). Most of these taxonomies have used the PDTB top-level classification and brought a number of adjustments to its sub-levels in order to account for all the specificities of their language. For example, in the Arabic version (Al-Saif and Markert, 2010), a *background* relation has been

added as a variety of expansion. This is therefore a case of specification with respect to the PDTB taxonomy. Conversely, the subtypes of *contrast* (opposition vs. juxtaposition) and *condition* (hypothetical, etc.) were removed from the Arabic taxonomy. This goes in the direction of a generalization of the taxonomy for these labels.

Another potential source of evidence for validating multilingual taxonomies comes from recent experiments with “translation spotting” of discourse connectives in parallel corpora (typically, Hansard or Europarl). Rather than annotate connectives in each monolingual part with PDTB-style labels, this approach aims at identifying (manually or automatically) the actual translation of each connective (Danlos and Roze, 2011; Popescu-Belis et al. 2012). This deals therefore only with explicit relations, not implicit ones. By clustering afterwards the observed translations according to their meaning and frequency, it is possible to derive labels which are less precise than the PDTB ones, but are still useful for applications such as machine translation (Meyer et al. 2011) or for translation studies (Cartoni et al., 2011).

Information from translation spotting can give a lower bound on the number of different meanings a connective can convey, which can be compared to the number of labels for that connective from a PDTB-style annotation, checking for any serious mismatch. For instance, if a connective is mainly annotated with one label, but is rendered in translation by two equally frequent target connectives, it is worth examining whether the sense label should not be *specified* any further.

Manual translation spotting has been performed on a large English/French section of the Europarl corpus with about 2,500 occurrences of nine connectives (Popescu-Belis et al. 2012). It is also currently being performed on English/German/Italian parallel fragments of Europarl within the same project. An experiment with automatic English/Arabic translation spotting, using word alignment software, is also ongoing for seven English connectives, illustrating ambiguity patterns (one vs. several preferred translations).

In what follows, we present two multilingual annotation experiments with explicit discourse relations in five European languages, with an

adaptation of the PDTB in between, using the two processes of specification and generalization introduced above.

4 Applying the PDTB taxonomy to a parallel corpus of five languages

4.1 Data and procedure

In order to compare and annotate connectives in five languages, a small parallel corpus made of four journalistic texts was gathered from the www.PressEurop.eu website. The size of the corpus was around 2,500 words per language. All four texts came from different European newspapers, and the source language was different in all of them. In the English version of the corpus, used as a pivot language, 54 tokens of connectives were identified, corresponding to 23 different connective types. Connectives were defined as lexical items encoding a coherence relation between two abstract objects, following Asher (1993). The criteria used to select tokens of connectives were similar to those applied in the PDTB project. However, only connectives that had been translated by a connective in a given language were annotated. This means that a slightly different subset of all the occurrences of English connectives was annotated in each case. The list of English connectives is given in Table 1.

after (1)	despite (1)	then (1)
after all (1)	for instance (1)	therefore (2)
and (7)	however (4)	though (2)
as (1)	if (2)	thus (2)
as long as (1)	in as much as (1)	when (4)
because (2)	meanwhile (1)	whereas (1)
before (1)	nevertheless (3)	while (1)
but (11)	so (1)	

Table 1. List of connective types from the English corpus with their token frequency.

Table 2 summarizes the number of connectives that have been inserted or removed in the target languages, with respect to the English texts. All these occurrences have therefore not been annotated.

In every language, the annotation task was performed independently by two annotators. The tokens of discourse connectives to be annotated were spotted on the English version of the

corpus by the authors. For every other language of the study, one annotator was asked to spot the translation equivalents. All tokens of connectives that had been translated in the target text by a connective were annotated with a discourse relation from the PDTB hierarchy by two annotators.

	French	German	Dutch	Italian
Nb. of additions	6	12	19	15
Nb. of removals	10	10	7	18
Total	16	22	26	33

Table 2. Differences in number of connectives between source and target texts.

All annotators were asked to use the definition of discourse relations provided in the PDTB annotation manual (The PDTB Research Group, 2007). As it was the case in the PDTB project, annotators were instructed to use tags from the most precise level from the hierarchy (third level) if they were confident about the relation or more generic relations in case of doubt. Annotators were also allowed to use two labels in two different cases: when they felt that the relation was ambiguous and that both tags applied; or when they felt that two tags had to be used in order to fully describe the meaning of the relation. In the first case, the two tags had to be linked with OR and in the second with AND.

4.2 Results

The inter-annotator agreement was computed from a monolingual and from a cross-linguistic perspective. The percentage of agreement for the two annotators working on the same language is reported in Table 3.

level	English	French	German	Dutch	Italian
1	98%	95%	95%	90%	94%
2	67%	69%	71%	60%	63%
3	44%	48%	51%	38%	42%

Table 3. Monolingual inter-annotator agreement.

Results from Table 3 indicate that the level of agreement is similar across languages. In every case, the agreement is very good at the first level (94% on average), medium at level 2 (66% on average) but poor at level 3 (44% on average).

By comparison, in the PDTB, the inter-annotator agreement was 92% at the top-most level and 77% at the third level of the hierarchy (Mitsalkaki *et al.*, 2008).

An analysis of cases of disagreement between the monolingual annotations reveals that similar problems occur in all languages. The problematic cases mostly concern the distinction between concession and contrast, for which the annotators agree in only 50% of the relations, when the ‘comparison’ tag is used. This agreement even drops to 40% on average at the third level (distinctions between opposition and juxtaposition and between expectation and contra-expectation). Moreover, for the relations tagged as ‘conditional’, the agreement for the third level tags is also only 40%. Taken together, these cases represent on average 87% of the disagreements at the third level of the hierarchy. Finally, the use of the so-called ‘pragmatic’ tags from the PDTB scheme was very problematic. An agreement on the use of this tag was reached only in 16% on the cases on average, and some annotators didn’t use it at all.

Cross-linguistic inter-annotator agreement is reported in Table 4.

level	English/ French	English/ German	English/ Dutch	English/ Italian
1	91%	90%	88%	85%
2	67%	66%	64%	58%
3	42%	51%	35%	35%

Table 4. Cross-linguistic inter-annotator agreement.

An analysis of cross-linguistic disagreements reveals two distinct phenomena. At the top level of the hierarchy, disagreements are always more numerous cross-linguistically than monolingually. These additional disagreements always correspond to meaning shifts due to translation. For example, the connective *when*, annotated with a temporal tag in English, was once translated by *alors que*, a connective annotated with a contrast tag by French-speaking annotators. Disagreements at the first level were systematically checked and discussed with annotators, with the conclusion that such cases of meaning shift occur on average in 10% of the cases in every language. This problem shows the limitations of using parallel corpora,

under the assumption that connectives are translation equivalents across languages. An annotation of comparable corpora, where equivalences are established based on the similarity of rhetorical relations, does not run into similar problems.

For lower levels of the hierarchy, differences in the annotation could not be related to changes in translation but rather to genuine disagreements between annotators regarding the interpretation of a given relation. For this reason, at these levels, disagreements are on average not significantly higher cross-linguistically than monolingually.

The first annotation experiment described above clearly indicated that the areas of disagreements were recurrent across annotators and languages. In order to reach a reliable annotation that could be applied cross-linguistically, some adjustments were made to the PDTB taxonomy.

5 Proposals for revisions to the PDTB taxonomy

First, through a generalization process, the sub-categories of conditional relations were removed because in all the languages of our study, all these uses were conveyed by a single connective (*if* in English, *si* in French, *als* in Dutch, etc.). For our objective to provide an accurate representation of the meaning of connectives enabling the definition of cross-linguistic equivalences in European languages, the second level *condition* tag is fine-grained enough.

Second, the categories labeled with the PDTB ‘pragmatic’ tag were redefined. In the PDTB taxonomy, the kind of examples grouped under this category was not always clearly defined and therefore was rather inconsistently applied by the annotators. For example, while a reference to epistemic uses is clearly made in the case of pragmatic causes, pragmatic conditions are simply defined as “used for instances of conditional constructions whose interpretation deviates from that of the semantics of ‘Condition’” (The PDTB Research Group, 2007: 31). In the revised version, the ‘pragmatic’ tag consistently includes all occurrences corresponding to speech-act and epistemic uses of connectives, as defined by Sweetser (1990).

Again, the rationale for this specification comes from differences in connectives. In many languages, content (non-pragmatic) and speech act and epistemic (pragmatic) relations are expressed by specific connectives (see Sanders and Stukker, 2012 for a cross-linguistic illustration in the causal domain). The pragmatic uses of connectives thus defined can occur for causal, conditional and concessive connectives. Therefore, for these tags, an additional annotation level has been specified to account for the pragmatic/non-pragmatic distinction. In the case of causals, this change involved the addition of a fourth level in the hierarchy. The addition of this level shows how certain semantic characteristics of relations occur across several categories, which leads to a systematic proposal (cf. Sanders et al., 1992).

<p>1. Temporal</p> <ul style="list-style-type: none"> - synchronous - asynchronous <ul style="list-style-type: none"> - precedence - succession <p>2. Contingency</p> <ul style="list-style-type: none"> - cause - reason <ul style="list-style-type: none"> - pragmatic - non-pragmatic - result <ul style="list-style-type: none"> - pragmatic - non-pragmatic - condition <ul style="list-style-type: none"> - pragmatic - non-pragmatic 	<p>3. Comparison</p> <ul style="list-style-type: none"> - contrast - concession <ul style="list-style-type: none"> - pragmatic - non-pragmatic - parallel <p>4. Expansion</p> <ul style="list-style-type: none"> - conjunction - instantiation - restatement <ul style="list-style-type: none"> - specification - equivalence - generalization - alternative - exception - list
---	---

Figure 1. Revised taxonomy based on the results of multilingual annotation.

Third, the *comparison* category was reorganized through a process of generalization. More specifically, the third level from the PDTB was removed, because it did not contribute to make additional distinctions between connectives. Furthermore, a ‘parallel’ tag was added, in order to account for the meaning of connectives such as *similarly*, which did not have a suitable tag in the PDTB taxonomy. All these changes lead to the revised taxonomy described in Figure 1. Similar adjustments were already proposed in some monolingual adaptations of the PDTB, notably in Arabic by Al-Saif and Markert (2010).

6 Annotation experiment with the revised taxonomy

A second corpus was gathered from the PressEurop website, including the same five languages used in the first experiment. This corpus, of about 8,500 words per language, contained in English 203 tokens of connectives corresponding to 36 different types (Table 5).

after (1)	given (that) (2)	since (1)
although (6)	however (7)	so (2)
and (50)	if (11)	that is why (1)
as (3)	in fact (1)	then (3)
as well as (1)	in order to (1)	therefore (3)
because (5)	in other words (1)	though (5)
before (4)	in short (1)	thus (2)
but (41)	in spite of (1)	well (1)
despite (6)	indeed (1)	when (7)
even if (4)	meanwhile (1)	whether (2)
for example (3)	now (2)	while (9)
for instance (1)	or (5)	yet (8)

Table 5. Connective types with token frequency.

In every language, the translation equivalents were spotted. The number of explicitly translated connectives ranged from 136 to 155. The important number of non-translated connectives provides further indication of the important volatility of these lexical items in translation. The rhetorical relations conveyed by explicit connectives were annotated with the revised taxonomy described in Figure 1. Results from the annotation task are reported in Table 6.

	English/ French	English/ German	English/ Dutch	English /Italian
level 1	94%	93%	88%	93%
level 2	85%	74%	75%	78%
level 3	75%	66%	69%	66%
level 4	66%	93%	62.5%	70%

Table 6. Cross-linguistic inter-annotator agreement.

These results confirm the validity of our second monolingual annotation experiment, with cross-linguistic data. The improvement of agreement scores with respect to the first experiment are significant, and the additional coverage of connective types did not reveal the need for additional relations or the existence of important differences between languages. This

experiment also confirmed that most disagreements at the first level of the taxonomy were due to meaning shifts in translation, as confirmed through manual checking and discussion with the annotators.

7 Conclusion

This paper is a first attempt towards a unified framework designed to relate connectives to one another over the languages. This existence of such a framework is a sorely needed resource for many domains such as applied linguistics, translation and language engineering. Such a resource is all the more necessary because existing multilingual resources such as bilingual dictionaries and contrastive grammars are insufficient to correctly describe them.

Yet, much work remains to be done to achieve this goal. Importantly, larger scale annotation experiments involving more languages and tokens for the annotation should be carried out. Another important step will be to test the granularity of the taxonomy by systematically comparing all tokens annotated with the same label, both monolingually and cross-linguistically, in order to ensure that they provide genuine semantic equivalences. In other words, the need for additional specifications should be systematically checked. Finally, another important step will be to include the implicit dimension in the cross-linguistic comparison of connectives. In some cases, the absence of connectives seems to be the preferred translation choice. A case in point is the French connective *en effet*, very frequently used to mark an elaboration, and most of the time not translated into English. Similar cases should be detected, and zero translations taken into account as possible translation equivalents.

Acknowledgments

This study was partially funded by the Swiss National Science Foundation through a grant awarded to the first author (SNSF Grant PA00P1_139613/1). The article was also supported by the COMTIS Swiss NSF Sinergia project (www.idiap.ch/comtis). The second author is Senior Research Associate with the Belgian National Science Foundation (FSR-

FNRS). The authors thank all the annotators who took part in the experiments.

References

- Al-Saif, A. and Markert K., 2010. "The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic". *Proceedings of the 7th Int. Conf. on Language Resources and Evaluation (LREC 2010)*, Marrakech, p.2046-2053.
- Asher, N. 1993. *Reference to Abstract Objects in Discourse*. Dordrecht: Kluwer.
- Bunt H., Prasad R. and Joshi A. 2012. "First steps towards an ISO standard for annotating discourse relations". *Proceedings of ISA-7 workshop (Interoperable Semantic Annotation) at LREC 2012*, Istanbul.
- Carston R. 2002. *Thoughts and Utterances. The Pragmatics of Explicit Communication*. Oxford: Blackwell.
- Cartoni, B. Zufferey, S. Meyer, T. Popescu-Belis, A. 2011. "How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives". *Proceedings of 4th Workshop on Building and Using Comparable Corpora*, Portland, OR.
- Danlos L. and Roze C. 2011. "Traduction (automatique) des connecteurs de discours". *Proceedings of TALN 2011*, 6 p., Montpellier.
- Danlos L., Antolin-Basso D., Braud C., and Roze C. 2012. "Vers le FDTB : French Discourse Tree Bank". *Proceedings of JEP-TALN-RECITAL 2012*, vol. 2, p. 471-478, Grenoble.
- Dixon, R.M. and Aikhenvald, A. (eds). 2009. *The Semantics of Clause Linking. A Cross-Linguistic Typology*. Oxford: Oxford University Press.
- Granger, S. and Tyson, S. 1996. "Connector Usage in English Essay Writing of Native and Non-Native EFL Speakers of English". *World Englishes* 15(1):19-29.
- Halverson, S. 2004. "Connectives as a Translation Problem". In *An International Encyclopedia of Translation Studies*, H.

- Kittel *et al.* (eds), 562-572. Berlin/New York: Walter de Gruyter.
- Huang, H.-H. and Chen, H.-H. 2011. "Chinese discourse relation recognition". *Proceedings of 5th International Joint Conference on Natural Language Processing*, Chiang Mai, Thailand, 1442-1446.
- Kolachina, S., Prasad, R., Sharma, D. and Joshi, A. 2012. "Evaluation of Discourse Relation Annotation in the Hindi Discourse Relation Bank". *Proceedings of LREC 2012 (8th International Conference on Language Resources and Evaluation)*, Istanbul.
- Meyer T., Popescu-Belis A., Zufferey S. and Cartoni B. 2011. "Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation". *Proceedings of SIGDIAL 2011 (12th annual SIGdial Meeting on Discourse and Dialogue)*, Portland, OR, p.194-203.
- Miltsakaki, E. Robaldo, L. Lee, A. and Joshi, A. 2008. "Sense Annotation in the Penn Discourse Treebank". *Lecture Notes in Computer Science* 4919: 275-286.
- Popescu-Belis A., Meyer T., Liyanapathirana J., Cartoni B. & Zufferey S. 2012. "Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns". *Proceedings of LREC 2012 (8th International Conference on Language Resources and Evaluation)*, Istanbul.
- Prasad, R. Dinesh, N. Lee, A. *et al.* (2008). "The Penn Discourse TreeBank 2.0". *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2961-2968.
- Sanders, T., Spooren, W. and Noordman, L. 1992. Towards a taxonomy of coherence relations. *Discourse Processes* 15: 1-35.
- Sanders, T. & Stukker, N. 2012. "Causal Connectives in Discourse: A Cross-Linguistic Perspective". *Journal of Pragmatics* 34(2): 131-137.
- Spooren, W. 1997. "The Processing of Underspecified Coherence Relations". *Discourse Processes* 24(1): 149-168.
- Spooren, W. and Degand, L. 2010. "Coding Coherence Relations: Reliability and Validity". *Corpus Linguistics and Linguistic Theory* 6 (2): 241-266.
- Sweetser, E. 1990. *From Etymology to Pragmatics*. Cambridge: Cambridge University Press.
- The PDTB Research Group. 2007. "The Penn Discourse Treebank 2.0 Annotation Manual". *IRCS Technical Reports Series*, 99 p.
- Webber B. and Joshi A. 2012. "Discourse Structure and Computation: Past, Present and Future". *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, p. 42-54, Jeju, Republic of Korea.
- Zeyrek D., Demirsahin I., Sevdik-Calli A., Balaban H.O., Yalcinkaya I., and Turan U.D. 2010. "The Annotation Scheme of the Turkish Discourse Bank and an Evaluation of Inconsistent Annotations". *Proceedings of the Fourth Linguistic Annotation Workshop at ACL 2010*, p. 282-289, Uppsala, Sweden.
- Zhou, Y. and Xue, N. 2012. "PDTB-style Discourse Annotation of Chinese Text". *Proceedings of ACL 2012 (The 50th Annual Meeting of the Association for Computational Linguistics)*, p.69-77, Jeju, Republic of Korea.
- Zikánová, S., Mladová, L. Mírovský, J. and Jínová, P. 2010. "Typical Cases of Annotators' Disagreement in Discourse Annotations in Prague Dependency Treebank". *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, p.2002-2006, Valetta.