Contrast-Ita Bank: A corpus for Italian Annotated with Discourse Contrast Relations

Anna Feltracco, Bernardo Magnini and Elisabetta Jezek

DOI: 10.4000/books.aaccademia.2407
Publisher: Accademia University Press
Place of publication: Torino
Year of publication: 2017
Published on OpenEdition Books: 19 April 2018
Serie: Collana dell'Associazione Italiana di Linguistica Computazionale
Electronic ISBN: 9788899982942

http://books.openedition.org

Electronic reference
Abstract

English. We present Contrast-Ita Bank, a corpus annotated with discourse contrast relations in Italian. We annotate both explicit and implicit contrast relations, following the schema proposed in the Penn Discourse Treebank. We provide and discuss quantitative data about the new resource.


1 Introduction

A relevant task in Natural Language Processing is the automatic identification of semantic relations between portions of text, such as textual entailment, text similarity, and temporal relation. In this contribution we focus on discourse contrast.

By discourse relation we mean a relation between two parts of a coherent sequence of sentences, propositions or speeches (i.e. discourse). We consider as discourse contrast: i) cases in which one of the two parts (henceforth arguments) is similar to the other in many aspects but different in one aspect for which they are compared, as in example (1), where both situations refer to a change in the price, but with different values; ii) cases in which one argument is denying an expectation that is triggered from the other argument, as in (2), where “going to the beach” denies the expectation that, since it is raining, one would stay home. Contrast in text can be conveyed explicitly, by mean of a lexical element (connective), as by while in (1) and although in (2), or implicitly as in (3).

(1) The price of x increased of 5%, while the price of y decreased of 2.3%.
(2) Although it was raining, we went to the beach.
(3) Mary passed the exam. John failed it.

We present Contrast-Ita Bank1, a corpus of Italian documents annotated with contrast, a very frequent relation in discourse. We aim to understand how frequent the contrast relation is in discourse, when it is expressed explicitly and implicitly, and which are the connectives that convey contrast. The final result of the annotation represents a first step toward a corpus of discourse relations for Italian, compatible with the Penn Discourse Treebank (PDTB) project (Prasad et al., 2007), the largest and the most used corpus annotated with discourse relations in the NLP field. A number of annotated corpora similar to the PDTB have been realised since its creation, for instance, the Prague Discourse TreeBank (Bejček et al., 2013), the Chinese Discourse TreeBank (Zhou and Xue, 2015), the Leeds Arabic Discourse TreeBank (Al-Saif and Markert, 2010)).2 For Italian, a similar attempt was proposed by Tonelli et al. (2010), which uses the PDTB scheme for the annotation of the LUNA conversational spoken dialogue corpus. The authors annotated 60 real dialogues in the domain of software/hardware troubleshooting. Another project for Italian inspired by the PDTB is proposed by Pareti and Prodanof (2010) and it is focused on the relation of attribution, i.e. “the relation of ownership between abstract objects and individuals or agents” (Prasad et al., 2007, p. 40).

Resources manually annotated with discourse relation have been used for instance for develop-

---

1https://hlt-nlp.fbk.eu/technologies/contrast-ita-bank
2Prasad et al. (2014) propose an overview of projects also mentioning resources for French, Turkish and Hindi.
The paper is structured as follows: Section 2 introduces the contrast relation; Section 3 describes the annotation guidelines; Section 4 presents the content of the resource and Section 5 discusses the inter annotator agreement.

2 The Contrast Relation

Discourse contrast has been described in various theories and annotation schema. In the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), contrast is defined as the relation between two spans of texts such that the situations presented in the two spans are: “(i) comprehended as the same in many respects, (ii) comprehended as differing in a few respects, and (iii) compared with respect to one or more of these differences” (Mann and Thompson, 1988). In the framework of RST, Carlson and Marcu (2001) propose a discourse relations corpus; in their schema, contrast is part of a broader class of relations called Contrast, together with concession, described as “characterised by a violated expectation”(Carlson and Marcu, 2001).

In the Segment Discourse Representation Theory framework, Asher and Lascarides (1993; 2003) define contrast as a relation that involves constituents that are structurally similar but semantically dissimilar. According to them, this relation includes cases of violation of expectation in which what can be inferred from one of the constituents of a relation is denied in the second constituent (Asher and Lascarides, 2003, p. 167).

The Penn Discourse Treebank schema (Prasad et al., 2007) proposes different senses of the connectives that provide a semantic description of the discourse relation they convey. These senses are annotated as sense tags. The sense tag CONTRAST applies to cases in which the two arguments of a relation “share a predicate or a property and a difference is highlighted with respect to the values assigned to the shared property”; the sense tag CONCESSION is used for cases in which “the highlighted differences are related to expectations raised by one argument which are then denied by the other” (Prasad et al., 2007).

We consider as contrast both what has been called formal contrast (Asher, 1993) and CONTRAST (Prasad et al., 2007) on the one hand (see Example (1) and (3)), and violation of expectation (Asher, 1993) or CONCESSION (Carlson and Marcu, 2001; Prasad et al., 2007) on the other hand (as in Example (2)).

3 Adopting the PDTB Schema

The Contrast-Ita Bank guidelines follow the PDTB 2.0 Annotation Manual (Prasad et al., 2007) and the recent proposal by Webber et al. (2016).

Following the PDTB 2.0, we annotate explicit relations (see Examples (1) and (2) above) by identifying the discourse connectives that trigger the relations and the respective arguments. We also annotate cases in which the relation is not marked by a connective and can be inferred between adjacent sentences. These cases include implicit relations, i.e. the relation is not lexically marked, as in Example (3), and alternatively lexicalized (altlex) relations, i.e. the relation is inferred by mean of another expression that is not a connective. By definition, these are cases where a discourse relation is inferred between adjacent sentences in absence of a connective, but where providing a suggestion of connective leads to redundancy in the expression of the relation (Prasad et al., 2007). For instance, in ‘She prepared a cake. The reason: it was his birthday.’, a cause relation is conveyed through ‘The reason:’; this relation is a case of Altlex, since ‘The reason:’ is not a connective, and providing a suggestion of connective (e.g. because) will lead to redundancy. Differently from the PDTB 2.0, we annotate implicit relations also among comma separated clauses and altlex among non adjacent sentences.

Specifically, our task involves: i) the annotation of the arguments of the relation (named Arg1 and Arg2, being Arg2, the argument in the clause that is syntactically bound to the connective, and Arg1, the other one); ii) the annotation of the connectives that convey contrast in the case of explicit relations, of the first token of Arg2 in the case of

---

1The task of identifying discourse relations in the form of a discourse connective taking two arguments is also called shallow discourse parsing and constituted a shared task of the CONLL conference in 2015 and 2016 (Xue et al., 2015).

2In the PDTB3.0 hierarchy (Webber et al., 2016), the two sense types belong to the class COMPARISON.

3See a similar example in (Prasad et al., 2007, p.7).
implicit relations, and of the expression that make us inferring the relation in the case of altlex relations; iii) the tagging of the sense of the relation. An example from the PDTB2.0 Manual (Prasad et al., 2007) is provided in (4), in which the connective appears underlined, Arg1 is in italics, and Arg2 is in bold.

(4) Most bond prices fell on concerns about this week’s new supply and disappointment that stock prices didn’t stage a sharp decline. Junk bond prices moved higher; however, (sense tag: Contrast)

Connectives. We followed the PDTB also for the definition of connectives that convey an explicit relation. They belong to three syntactic classes: (i) subordinating conjunctions (e.g. when, because); (ii) coordinating conjunctions (e.g. and, or, but); (iii) discourse adverbials, including both adverbs (e.g. however, instead), and prepositional phrases (e.g. on the other hand, as a result).

Arguments. According to the PDTB, relations are annotated when they are connecting “two abstract objects such as events, states, and propositions (Asher, 1993)” (Prasad et al., 2007), that are realised mostly as clauses, nominalisations, or anaphoric expressions. We follow the same guidelines, including conjoined VPs, as proposed by Webber et al. (2016).\(^6\) We also adopt the Minimality Principle, according to which “only as many clauses and/or sentences should be included in an argument selection as are minimally required and sufficient for the interpretation of the relation”(Prasad et al., 2007). This means that there is no constrain on the length of an argument or that more than a sentence can be annotated (i.e. punctuation is generally not a limiting constrain).

Senses of relations. We consider a broad semantic definition of contrast, corresponding to the PDTB sense tags CONTRAST and CONCESSION. Specifically, we follow the PDTB 3.0 schema (Webber et al., 2016) in which CONCESSION has two subtypes, depending on which argument creates the expectation and which one denies it: if Arg2 creates an expectation that Arg1 denies, the proper tag is CONCESSION_Arg1.as.denier; conversely, when Arg1 creates an expectation that Arg2 denies, the tag that needs to be used is CONCESSION_Arg2.as.denier. In line with the PDTB 2.0 we allow the annotation of more than one sense for a connective and, thus, the possibility of marking e.g. both CONTRAST and CONCESSION_Arg1.as.denier. Table 1 summarises the definition of the tags.

<table>
<thead>
<tr>
<th>Relation and Definition in the PDTB</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTRAST → if the two Args share a predicate or a property and the difference between the two situations (in the Args) is highlighted with respect to the values assigned to the property.</td>
<td></td>
</tr>
<tr>
<td>CONCESSION → if one argument which are then denied by the other.</td>
<td></td>
</tr>
<tr>
<td>Arg1.as.denier if Arg1 denies expectation</td>
<td></td>
</tr>
<tr>
<td>Arg2.as.denier if Arg2 denies expectation</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: CONTRAST and CONCESSION in the PDTB 3.0 (Webber et al., 2016).

4 Contrast-Ita Bank

Contrast-Ita Bank is based on a corpus of 169 news stories selected from Ita-TimeBank (Caselli et al., 2011), for a total of 65,053 tokens (average length = about 385 tokens per document).\(^3\) For the annotation we used the CAT tool (Bartalesi Lenzi et al., 2012). The annotation was carried by one expert annotator in about two weeks.

We annotated explicit, implicit and altlex relations of contrast for a total of 372 relations (average 2.16 per document). Table 2 reports the data of the annotation. Explicit relations are the most common and correspond to 91% of all the relations. We register a maximum number of 15 explicit relations in one document and an average of 2 relations per document. Implicit relations are less frequent and occur 15 times inter-sententially and 9 times infra-sententially, for a total of 24 annotations. This is different from the PDTB2.0, in which the ratio between explicit and implicit for what concerns CONTRAST and COMPARISON, and their subtypes, is about 0.45, while in Contrast-Ita Bank is ten time less. This might be due to the fact that in Contrast-Ita Bank annotators were asked to mark contrast, and it is possible that they simply fail to capture implicit relations, while in the PDTB2.0 annotators were asked to mark also cases where no relation can be inferred between adjacent sentences, thus analysing in detail if a relation appears between every pair of sentences. Altlex relations are rarer: in Contrast-Ita Bank...
Table 2: Contrast relations in Contrast-Ita Bank.

<table>
<thead>
<tr>
<th>Connective</th>
<th>Explicit</th>
<th>Implicit</th>
<th>AltLex</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTRAST</td>
<td>87</td>
<td>12</td>
<td>3</td>
<td>102</td>
</tr>
<tr>
<td>CONC.Arg1-denier</td>
<td>21</td>
<td>0</td>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>CONC.Arg2-denier</td>
<td>201</td>
<td>8</td>
<td>3</td>
<td>212</td>
</tr>
<tr>
<td>Double relations</td>
<td>32</td>
<td>4</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>341</td>
<td>24</td>
<td>7</td>
<td>372</td>
</tr>
</tbody>
</table>

Table 2 also shows that the per token density of contrast in the corpus is 0.0056, similar to the PDTB (i.e. 0.0072).

The most frequent sense tag is CONCESSION. Arg2-as-denier (i.e. when Arg2 denies an expectation that rises from Arg1), which covers about 56% of the cases. CONTRAST covers almost a quarter of the cases and the two relations have been annotated together 32 times (out of the total 36 cases of double annotation). CONCESSION.Arg1-as-denier is far less frequent both as single type as with other relations, and has been annotated less than 10% of the cases. This subtype is associated to a limited set of connectives: despite the list of connectives in Contrast-Ita Bank consists of 19 connectives (see Table 3), 7 of them (e.g. nonostante) signal CONCESSION.Arg1-as-denier all the times.

Not surprisingly, ma accounts for almost half of the cases (the equivalent but is also the most used for these senses in the PDTB 2.0), and invece, mentre, però for about a 10%. Table 3 shows that, as it happens for content words, the most frequent connectives are the most polysemous ones.

5 Inter Annotator Agreement

We computed the agreement (IAA) between two annotators on 18 documents (10.6% of the whole corpus), which followed the same written guidelines. Data are reported in Table 4.

---

This is also the rarest type in the PDTB 2.0, among the three considered here.

It is possible that contrast is more frequent in corpora of other domains, such as in documents reporting debates in which people contrast their opinions. However, with the idea of maximising the compatibility with the PDTB, we annotated contrast on a corpus of news.

---

The Dice’s coefficient measures how similar two sets are or if one marked ma or if one marked se and the other anche se to signal the presence of a contrast relation. We calculated the final score adopting the Dice’s coefficient (Rijsbergen, 1979). The result is that annotators agree in 37 cases (Dice 0.68). We consider this result reasonable given the difficulty of the task which has not to be underestimated. To identify contrast relation in a document means to distinguish cases in which a lexical element is playing the role of connective of contrast or it is not, and also to identify implicit relations that by definition are not marked in the text. In order to understand the motivations of these discrepancies, we have adopted a reconciliation strategy among annotators in which they were asked to motivate their choices with the possibility of revising them. After the reconciliation dis-
Discussion 16 cases were reconciliated and the Dice value increased to 0.84.

In other cases disagreement remained. These mainly include cases in which both annotators recognized a discourse relation but one interpreted the relation to be of contrast, while the other did not. In many cases, these relations are conveyed by the coordinating conjunction ‘e’. We report an example in which one annotator recognized a contrast; while the other considered the arguments as non-contrasting parts of a description.

(5) [...] sono portatori sani di Talassemia Mayor e il loro bambino, Luca, cinque anni, è talassemico.\textsuperscript{11} [doc:5402]

CONTRAST vs NON-MARKED

Agreement on connectives identification is calculated considering if both annotators agree on recognising the same explicit relation and the same exact span of text to be a connective (thus excluding cases of altlex and implicit). In these terms, cases of agreement for connectives identification are a subset of cases of agreement already captured by the relation identification. The resulting agreement is 0.68 (Dice’s coefficient).

For the 37 cases of agreement on relation identification, we calculated the IAA on the span of arguments in two ways. In the exact match mode, we have agreement if the two annotators consider the exact span of text as Arg1 or Arg2 for the same relation; in the relaxed match mode, we consider agreement if the text span identified by the annotators matches at least for its 50%. Agreement in the exact match for Arg1 is 0.51 and for Arg2 is 0.70; in the relaxed match mode is 0.89 for Arg1 and 0.91 for Arg2. We expected the exact match agreement difficult to reach. In fact, as described in Section 3, we adopt the Minimality Principle for the annotation of the arguments. The selection of the arguments span thus relies significantly on the interpretation of the annotators and cases in which there is no exact match can be frequent.

Agreement in identifying CONTRAST and CONCESSION (sense type) is calculated counting 1 point if annotators agree to assign (or not) the same tag(s), 0.5 if one chooses a tag and the other both, 0 for total disagreement. IAA is obtained summing the points for each annotation and dividing by the total of 37 relations that both annotators identified. Agreement for sense type is

\begin{table}
\centering
\begin{tabular}{|l|c|}
\hline
\textbf{IAA on:} & \\
\hline
relation identification & 0.68 \\
relation identification - post reconciliation & 0.84 \\
connectives identification - explicit & 0.68 \\
arguments span - exact match (Arg1; Arg2) & 0.51; 0.70 \\
arguments span - relaxed match (Arg1; Arg2) & 0.89; 0.91 \\
sense type: CONTRAST - CONCESSION & 0.73 \\
sense subtype: Arg1.as.denier - Arg2.as.denier & 0.9 \\
\hline
\end{tabular}
\caption{InterAnnotator Agreement.}
\end{table}

0.73, showing that recognising the type of contrast can be a controversial decision among annotators. However, we believe that this result is fair, considering that the annotation regards non mutually exclusive types of the same class.

Finally, when there is agreement on CONCESSION, we applied the same formula to calculate IAA between CONCESSION subtypes: Arg1.as.denier - Arg2.as.denier: agreement is 0.9. Specifically, annotators agree in 10 cases to mark CONCESSION but in one case they disagree over the direction of the relation.\textsuperscript{12}

Overall, the IAA highlights that the main difficulties of annotating contrast concern: the relation identification, especially for implicit and altlex relations; the extent of the arguments: the two annotators frequently do not mark exactly the same tokens but it is very likely that their annotations match at least for their 50%; sense type: one annotator tends to annotate also the CONCESSION_ARG2.as.denier when marking CONTRAST, while the other annotator does not.

6 Conclusion and Further Work

We presented Contrast-Ita Bank, a corpus annotated with discourse contrast relations in Italian. Following the PDTB annotation schema, we annotated \textit{explicit}, \textit{implicit} and \textit{altlex} relations of contrast. We also present the list of connectives that convey contrast in the corpus. The new resource can be integrated with LICO, the Lexicon of Italian Connectives (Feltracco et al., 2016), validating the list of connectives and adding examples from corpus to the connectives. Contrast-Ita Bank

\textsuperscript{11}Eng.: [...] they are carrier of Talassemia Mayor and their son, Luca, five years old, is thallassemic.

\textsuperscript{12}For the argument identification in the PDTB 2.0, Prasad et al. (2008) report an agreement of 90.2% for explicit relation and 85.1% for implicit (we do not calculate the value considering this granularity); when relaxing the match to partial overlap, the two values increase to 94.5% and to 85.1%. Additionally, authors report an agreement of 94% for sense class, of 84% for sense type, and of 80% for the subtype level.
is distributed under a CC-BY-NC 4.0 licence.

References


