# Annotation of Discourse Relations for Conversational Spoken Dialogs

**Sara Tonelli**[1], **Giuseppe Riccardi**[2], **Rashmi Prasad**[3], **Aravind Joshi**[3]

[1]Dept. of Language Sciences, University of Venice, [2]DISI, University of Trento, [3]IRCS, University of Pennsylvania
satonelli@fbk.eu; riccardi@disi.unitn.it; {rjprasad, joshi}@seas.upenn.edu

**Abstract**

In this paper, we make a qualitative and quantitative analysis of discourse relations within the LUNA conversational spoken dialog corpus. In particular, we describe the adaptation of the Penn Discourse Treebank (PDTB) annotation scheme to the LUNA dialogs. We discuss similarities and differences between our approach and the PDTB paradigm and point out the peculiarities of spontaneous dialogs w.r.t. written text, which motivated some changes in the sense hierarchy. Then, we present corpus statistics about the discourse relations within a representative set of annotated dialogs.

## 1. Introduction

The study of applications and approaches able to capture syntactic and semantic relations beyond the sentence level has deserved increasing attention by the NLP community. For many NLP tasks such as text summarization, language generation and dialog management, the information acquired at the sentence level is clearly insufficient and systematic work in corpus analysis at the discourse level is required.

One of the main current efforts in this direction is the creation of the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), a corpus of English texts from the Wall Street Journal where relations between abstract objects in discourse such as propositions and eventualities are annotated. This paradigm, which was first developed for English texts, was then applied in a similar way to other languages such as Chinese (Xue, 2005), Turkish (Zeyrek and Webber, 2008), Czech (Mladová et al., 2008) and Hindi (Oza et al., 2009).

With this work, we address a two-fold issue: on the one hand, we apply for the first time the PDTB paradigm to Italian texts and analyze the usage of the most common connectives, comparing them to their English translation equivalents, if available. On the other hand, we investigate the applicability of the PDTB annotation scheme to spontaneous dialogs and propose a possible revision of the sense hierarchy, taking into account pragmatic aspects of conversational speech.

This work is structured as follows: in Section 2. we introduce the LUNA corpus and the three-layered annotation protocol devised in the LUNA project. In Section 3. we present the PDTB annotation framework, with details about the relation types and the sense labels. In Section 4. we present the criteria followed in annotating the LUNA corpus with discourse relations. In particular, we define explicit and implicit relations, we describe the argument selection step and then discuss the new sense hierarchy proposed. Then, in Section 5., we further describe the annotated data from a quantitative point of view and we report some statistics about the most frequent connectives and the most frequent senses identified both in implicit and explicit relations. We eventually draw some conclusions and describe future research directions in Section 6.

## 2. The LUNA corpus

In the context of the European project LUNA (Language UNderstanding in multilinguAl communication systems)[1], a corpus of spoken dialogs in Italian, French and Polish was acquired to study new solutions for Spoken Dialog Systems, specifically to enhance real-time understanding of spontaneous speech in advanced telecom services.

The project focused on different objectives, namely the language and semantic modeling of speech, and the automatic learning and the multilingual portability of spoken language understanding components.

In this framework, a considerable part of the work about semantic modeling of dialogs consisted in the multi-layered annotation of a corpus of Italian spontaneous speech recorded in the help-desk facility of the Consortium for Information Systems of Piedmont Region. The corpus contains 1000 equally partitioned Human-Human (HH) and Human-Machine (HM) dialogs. The former are real conversations about software/hardware troubleshooting, while the latter are dialogs where an operator acting as Wizard of Oz reacts to the caller's requests following one of ten possible scenarios.

The above data is organized in transcriptions and annotations of speech based on a new multi-level protocol studied specifically within the project, i.e. the annotation levels of words, turns, dialog acts, attribute-value pairs and predicate argument structures (Dinarelli et al., 2009). The dialogs are first recorded as audio files and then segmented at turn level and semi-automatically transcribed. Then, they are further segmented by hand at utterance level[2] and are annotated at three parallel semantic levels:

[1]EU FP6 contract No. 33549, http://www.ist-luna.eu/

[2]The time interval of each speaker's activity is defined as a *turn*, which is included between two pauses in the speech flow. *Utterances* are complex semantic entities that usually represent the annotation unit for dialog acts. Their relation to speaker turns is not one-to-one, because in most cases a single turn contains multiple utterances, and sometimes utterances can span more than one turn.

i) domain attribute annotation, based on a pre-definite domain ontology, where concepts and their relations are specified

ii) dialog act annotation, following the taxonomy described in (Quarteroni et al., 2008)

iii) annotation of predicate-argument structure, based on a domain adaptation of the FrameNet paradigm (Tonelli and Riccardi, 2010)

Despite the rich semantic annotation provided, we believe that, in the design of conversational systems, semantic information identified within the turn boundaries is not enough to drive the dialog strategy, typically organized across different dialog turns. For this reason, we decided to add a further annotation layer providing information to a much greater extent than just across turn boundaries, namely across all utterances, both within and across turn boundaries. This annotation was aimed at identifying *discourse relations*, such as "causal", "contrastive", "temporal" relations, between utterance transcriptions. To this purpose, we adopted the approach followed in the Penn Discourse Treebank (PDTB) project, where discourse relations are treated as discourse-level predicates taking two abstract objects (AOs), such as eventualities and propositions, as arguments. Such approach is particularly suited to the annotation of dialogs because it is theory-neutral, i.e. no theory-driven high-level structures are inferred from low level annotations of relations, so that it can be easily applied to different frameworks, including spontaneous dialogs.
Details about the PDTB project are reported in the following section.

## 3. The Penn Discourse Treebank

The Penn Discourse Treebank (Prasad et al., 2008) is a resource built on top of the Wall Street Journal (WSJ) (Marcus et al., 1993) consisting of a million words annotated with discourse relations. Discourse connectives are seen as discourse predicates taking two text spans as arguments, that correspond to propositions, events and states.
Discourse relations are realized in three ways: (a) as explicit connectives, (b) as alternative lexicalizations (AltLex) and (c) as implicit relations. Explicit connectives belong primarily to a few well-defined syntactic classes, while alternative lexicalizations are generally non-connective phrases used to express discourse relations, such that the insertion of an explicit connective would lead to redundancy. Implicit connectives, instead, express implicit discourse relations inferred between adjacent sentences, for which the annotator can insert a connective to express the inferred relation.
The abstract objects involved in a discourse relation are called `Arg1` and `Arg2` according to syntactic criteria and each relation can take two and only two arguments. We report below three example sentences from the PDTB showing respectively explicit, AltLex and implicit relations. `Arg1` is reported in italics, `Arg2` appears in bold and discourse relations are underlined. Note that the connective "So" in (c) is not present in the text, but is manually indicated by the annotator; it is written in capitals in order to distinguish it from connectives that appear in the text and are therefore explicit.
The above mentioned notation conventions will be applied to all examples reported in this paper.

**(a)** Explicit: *The federal government suspended sales of U.S. savings bonds* because **Congress hasn't lifted the ceiling on government debt.**

**(b)** AltLex: *Ms. Bartlett's previous work, which earned her an international reputation in the non-horticultural art world, often took gardens as its nominal subject.* **Mayhap this metaphorical connection made the BPC Fine Arts Committee think she had a literal green thumb**.

**(c)** Implicit: *The projects already under construction will increase Las Vegas's supply of hotel rooms by 11,795, or nearly 20%, to 75,500.* SO **By a rule of thumb of 1.5 new jobs for each new hotel room, Clark County will have nearly 18,000 new jobs**.

In the PDTB, a fourth relation type called EntRel was identified when no discourse relation (such as a causal or contrastive relation) was inferred between adjacent sentences, and when the connection between them involved only an entity-based coherence.
Each discourse relation is assigned a sense label based on a three-layered hierarchy of senses. The top-level, or *class level*, includes four major semantic classes, namely TEMPORAL, CONTINGENCY, COMPARISON and EXPANSION. For each class, a more fine-grained classification has been specified at *type* level. For instance, the relation in example (a) belongs to the CONTINGENCY class and the *Cause* type. A further level of *subtype* has been introduced to specify the semantic contribution of each argument. *Cause*, for instance, comprises the *reason* and the *result* subtypes. The former applies when the situation described in `Arg2` is the cause of the situation in `Arg1`, like in example (a), while the latter indicates that the situation in `Arg2` is the result of the situation in `Arg1`.
The annotation scheme was developed and refined by the PDTB group in a bottom-up fashion, following a lexically grounded approach to annotation. For the LUNA corpus, we adopt the same annotation strategy, though some major changes in the argument selection and the sense hierarchy are required to cope with the specific features of dialogs.

## 4. Discourse annotation and the LUNA corpus

In order to carry out a preliminary investigation of the connections between clauses and turns in a dialog, we annotated with discourse relations 60 HH dialogs from the LUNA corpus. We focused on human-human dialogs because we wanted to capture the complexity of real conversations and not follow the predefined structure of human-machine dialogs.
Annotation was carried out on raw text, regardless of the existing information that had already been encoded (see the annotation layers described in Section 2.). Overlaps were

recorded in a separate file and were not visible in the raw dialogs. This means that overlapping turns were just displayed in sequence and the annotator had to reconstruct the turn span following his intuition and the content of text segments.

### 4.1. Argument selection strategy

The basic intuition for argument selection remained the same as in the PDTB: for each discourse relation, we identified in the LUNA corpus two arguments, `Arg1` and `Arg2`, assuming that each relation can have two and only two arguments. Text span selection followed the "minimality principle", i.e. only the text string minimally necessary to interpret the relation was selected for each argument.

A relevant adjustment we had to introduce in the argument selection was that we could not limit annotation of implicit relations to adjacent sentences or turns as in the PDTB, because discourse in dialogs is much more fragmented than in prose and there are a lot of interruptions, disfluencies, etc. Keeping the adjacency criterion would have implied missing a lot of implicit relations, so we just suggested that all implicit relations should be identified in the text. An example of an implicit relation between two non-contiguous arguments is reported in (d). We mark with index 1 and 2 the utterances expressed respectively by Speaker 1 and 2.

**(d)** Implicit: "*Anche questo non è attivo*"$_1$ "quindi possiamo contattarla al"$_2$ "PERÒ **sto aspettando che me l'attivino**"$_1$

"*This is not active either*"$_1$ "So we can contact you at"$_2$ "BUT **I'm waiting for it to be activated**"$_1$

The two arguments are part of the same turn, even if they are not adjacent, while the sentence "So we can contact you at" is clearly an interruption inserted in the dialogs by a different speaker.

### 4.2. Relation types

As in the PDTB, we annotated in the LUNA corpus four relation types: *Explicit* discourse connectives, *Implicit* relations, *AltLex* and *EntRel*. Besides, we introduced the *Interruption* label for the cases in which the speaker has been interrupted while uttering a sentence and therefore he could just express one complete argument.

Explicit connectives are considered to build a *closed class*, drawn from three grammatical classes: i) subordinating conjunctions: ii) coordinating conjunctions iii) ADVP and PP adverbials

Not all tokens of words and phrases that can serve as Explicit connectives actually do so. In some cases, which are very frequent in spontaneous speech, they do not denote relations between two abstract objects, thus they have not been annotated as discourse connectives. In particular, there is a group of words including adverbials and connectives that are commonly defined as *discourse markers* (Schiffrin, 1987) or *phatic connectives* (Bazzanella, 1990) such as "cioè" (well), "allora" (so), etc. These words have not been annotated when they are used to signal the organizational or focus structure of the discourse and underline the interactive structure of the conversation, rather than relate AOs. Note that most of such words appear in the dialogs also as proper connectives. For a comparison between the use of "allora" (so) as discourse marker and as connective, see the examples below. In example (e), "Allora" is clearly a turn-taking device, while in example (f) it connects two turns and introduces a causal inference drawn by Speaker 2.

**(e)** "Allora vediamo un po' ecco qua"$_1$
*"So let's see here it is"*$_1$

**(f)** "*In questo momento il palazzo non è collegato*"$_1$ "Allora **è meglio collegarlo**"$_2$

"*In this moment the building is not connected*"$_1$ "So **we'd better connect it**"$_2$

There are also other cases in which some words and phrases that can serve as explicit connectives serve other functions, such as to relate non-AO entities, and are not annotated as discourse connectives. This is the case for example of "e" (and) conjoining two noun phrases, or "quindi" (so/namely) modifying a noun phrase.

As for implicit connectives, the identification and annotation of a discourse relation is the same as in the PDTB: in order to capture relations between abstract objects that are not explicitly realized in the text, annotators have to first identify the arguments involved in the relations and then insert a connective expression that best expresses the inferred relations. Insertable connectives are drawn primarily from the set of explicit connectives, but annotators are free to select alternative expressions as well. Also, combinations of connectives are allowed. An example of an implicit relation is reported in example (d), with the connective "però" (but) manually specified by the annotator.

As for alternative lexicalizations or AltLex, several examples are present in the LUNA corpus. One of them is reported below:

**(g)** "**Forse lei prima tentava di accedere con le iniziali del nome e del cognome**"$_1$ "Ecco perché *non riuscivamo*"$_2$

"**Maybe you were trying to login using the initials of your name and surname"**$_1$ "That's why *we couldn't*"$_2$

Example (g) is a typical case of an alternative lexicalization because the relation between the arguments is conveyed by a non-connective expression ("Ecco perchè") having two parts, one referring to the relation ("perchè") and the other referring anaphorically to the previous argument ("Ecco"). Similar cases are "Per questo motivo" (For this reason), "Nonostante questo" (Despite this), "Dopo questo evento" (After this event), and so on. As shown in example (g), we cannot classify such relations as implicit, because the insertion of a connective between the arguments would be redundant.

In order to make the causative relation more explicit, we could reformulate the two turns in (g) as:

"We couldn't access"$_2$ "because maybe you were trying to login using the initials of your name and surname"$_1$.

### 4.3. Sense Hierarchy

As in the PDTB, every discourse relation found in the LUNA corpus was classified with a sense label describing the semantics of the relation. Also, the LUNA senses

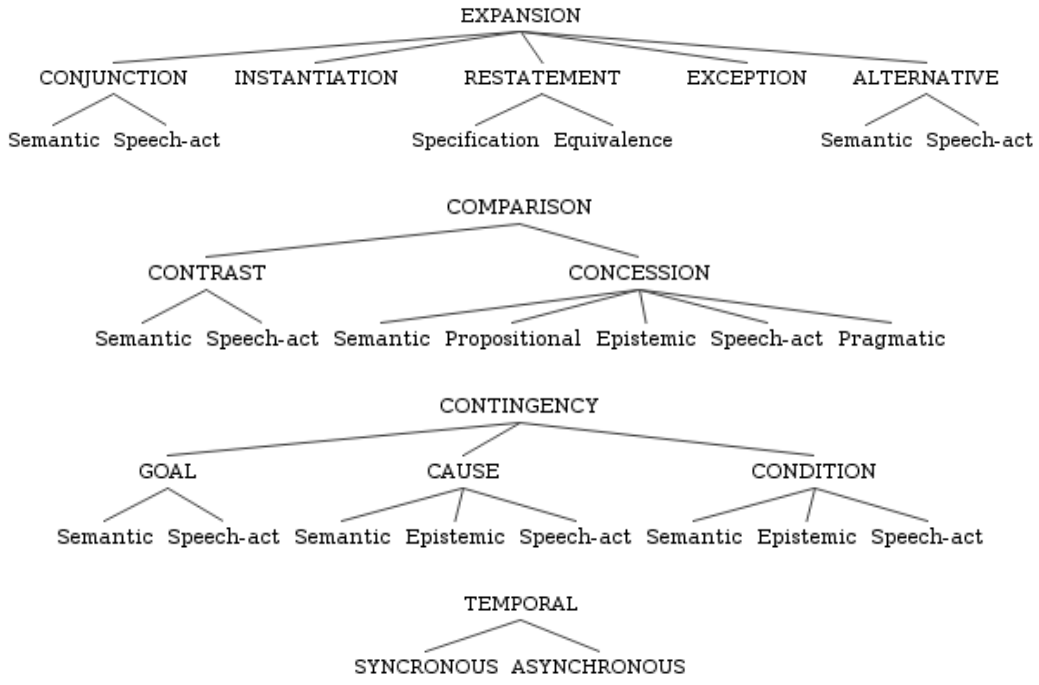


Figure 1: Sense Hierarchy in the LUNA discourse annotation

follow a hierarchical three-layered classification, shown in Figure 1.

At the top level, or class level, the four major semantic classes remain the same as in PDTB. The TEMPORAL class refers to relations in which the situations described in the arguments are related temporally. CONTINGENCY describes relations in which one situation is causally influenced by the other. COMPARISON applies when the discourse relation between `Arg1` and `Arg2` highlights prominent differences between the two situations. Finally the EXPANSION class refers to the relations that expand the discourse and move its exposition forward.

The second annotation level, i.e. the *type*, defined to describe in a more fine-grained way the semantics of the classes, presents a few differences with respect to the reference hierarchy of the PDTB. While in the PDTB the CONTINGENCY class was divided into two types, namely *Cause* and *Condition*, we have added a third type, *Goal*, which applies to relations where the situation described in one of the arguments is the goal of the situation described in the other argument. We define as `Arg1` the situation that enables / is aimed at the achievement of the goal, and `Arg2` the goal. An example is reported in (h):

**(h)** "*Chiamo subito l'help-desk* così **viene un tecnico a fare un controllo**"
*"I call the assistance right now* so that **a technician will come and check it**"

Another adjustment we introduced at type level was the elimination of the *List* type in the EXPANSION class. The *List* type in the PDTB was applied when `Arg1` and `Arg2` are members of a list defined in the prior discourse. In the LUNA corpus no relations were found that could be described using the *List* label, probably because discourse in conversational speech is less structured than in the WSJ prose.

The most relevant differences between the PDTB and the LUNA hierarchy involve the third classification level, i.e. the *subtype*. These modifications were partly inspired by the classification proposed in the Hindi Discourse Relation Bank (Oza et al., 2009). Although the majority of the relations in the PDTB express a *semantic* meaning between the arguments, a few relations are introduced to capture interpretations that do not directly involve the situations described by `Arg1` and `Arg2` but rather the *intention* of the author or an *epistemic inference* that needs to drawn from the text. Such cases were generally treated as a *pragmatic inference* and labeled as *pragmatic* (for example as "Pragmatic contrast" or "Pragmatic condition"). An example of "Pragmatic cause" from the PDTB is reported in (i). There is no causal influence between the two situations, since `Arg2` does not express the *cause* of the situation in `Arg1` but rather of why the author believes `Arg1` to be true.

**(i)** "*Mrs Yeargin is lying* because **they found students in an advanced class a year earlier who said she gave them similar help**"

While the pragmatic interpretation of discourse relations is not frequent in the PDTB, the speaker's intention, inferences, implicit connections in a dialogue are fundamental to understand the dialogue structure. For this reason, the pragmatic label was not sufficient to specify the non-semantic interpretation of connectives in the LUNA corpus and a further refinement was required. We introduced a more fine-grained classification of the pragmatic senses labeled in the PDTB, because the meaning of connectives in dialogs relate more to the epistemic or conversational domains than

to the content or semantic domain (Sweetser, 1990).
Furthermore, in contrast to the PDTB, where the pragmatic senses are specified at the type level, we introduce them at the subtype level, distinguishing them from the semantic senses, as shown in Fig. 1. Whenever the pragmatic senses are available for the relation, the corresponding type level sense is distinguished at the subtype level into its semantic and pragmatic senses. In general, we admit two kinds of pragmatic senses, i.e. *epistemic* and *speech-act*. Only for *Concessions*, we introduced two more subtypes, i.e. the proper *pragmatic* and the *propositional* one, because the available ones could not capture all examples of concession found in the corpus[3].
The epistemic label is assigned when the speaker's opinion, belief, interpretation is involved in the relation, while the speech-act subtype applies when the relation concerns the speech-act level and not strictly the meaning of the argument(s) (Berretta, 1984). Two examples of epistemic and speech-act type of causal relation are reported resp. in (l) and (m):

**(l)** "*Ho il PC che presumibilmente non funziona da*"$_1$ "sì"$_2$ "*stamattina* perché ho acceso **dà un segnale sul video tipo televisore senza antenna**"$_1$.

"*My PC hasn't presumably been working since*"$_1$ "yes"$_2$ "*this morning* because I switched it on **it shows a signal on the video like a TV without antenna**"$_1$.

**(m)** "*Avrei bisogno di sapere qualcosa al riguardo della richiesta numero centosessantaquattro diciassette* perché **avevate mandato la mail**"$_1$.

"*I would like to know something about my request number one hundred sixty-four seventeen* because **you had sent me an e-mail**"$_1$.

In (l), the fact expressed in `Arg2` (in bold) causes the fact that *the speaker believes* the content of `Arg1` (in italics). In other words, we classify this relationship as *epistemic* because `Arg1` expresses a speaker's belief or conclusion that is based on an observation or justification displayed in `Arg2`.
In (m), `Arg2` explains why the speaker is *asking* the indirect question in `Arg1`. We can say that the causal relation does not involve the semantics of the two events described in the two arguments but rather the *speech-act* level of `Arg1` and the reason motivating the speech-act (expressed in `Arg2`).
While we introduced new labels at subtype level, we eliminated some other subtype labels of the PTBD, many of which were just expressing a variation in the order of the arguments. For example, in the PDTB the *Cause* type is divided into the *reason* and *result* subtypes. In the former case, the situation described in `Arg2` is the cause of the situation in `Arg1`, while it is the contrary for the *result* subtype. In all cases, the naming convention for `Arg1` and `Arg2` is *syntactically driven*, in that `Arg2` always corresponds to the argument with which the connective was syntactically associated while the other argument is expressed in `Arg1`.

[3]The distinctions introduced for concessions are still under discussion and will probably undergo a further revision.

In the LUNA corpus, instead, the argument identification is *semantically driven*, i.e. every argument bears a sense-specific semantic role regardless of its position in the relation. In this way, we could merge the *reason* and *result* subtypes under the *cause* type, assigning the `Arg2` label to the situation that causes the event expressed in `Arg1`. According to this classification, both examples in (n) and (m) report a relation classified as (semantic) cause. `Arg1` (in italics) precedes `Arg2` (in bold) in the first example, while the order is inverted in the second example.

**(n)** "*Hanno di nuovo chiamato* perché **c'erano ancora dei problemi**".
"*They called again* because **there were still problems**".

**(o)** "**La fotocopiatrice si inceppa sempre** quindi *abbiamo dovuto togliere i fogli*."
"**The photocopier always jams** so *we had to extract the paper*".

In dialogs, a clause, a sentence or a turn is often the exact repetition of a previous utterance or part of it due to the interactive nature of spontaneous conversations. We decided to annotate such cases introducing the *Repetition* label because repetitions in LUNA were very frequently used by the speakers as a device to connect different turns. We consider *Repetitions* as a particular kind of implicit relations, which however do not require any connective to be specified.
We report an example of Repetition in (q), where Speaker 2 repeats part of the utterance by Speaker 1:

**(q)** "Allora *ho tolto le eccezioni* funziona"$_1$ "**hai tolto**"$_2$ "riprova"$_1$ "**le eccezioni**"$_2$

"So *I disabled the exceptions* it works"$_1$ "**You disabled**"$_2$ "Try it again"$_1$ "**the exceptions**"$_2$

The above example shows also that it is not always easy to understand who's speaking and to identify the relations between utterances in a dialog. In this case, `Arg2` (in bold) is discontinuous because "Try it again" is overlapping part of it. But also "*You disabled*" was uttered to interrupt the first turn. The example also shows that arguments, for instance `Arg1` (in italics), do not necessarily coincide with turns, but rather that they mostly include part of them.

## 5. Corpus analysis

We report in Table 1 some statistics over the Human-human dialogs annotated so far. For the sake of simplification, *Implicit* relations also include *Repetitions*.
In this corpus, the number of annotated relations is less than half of the number of turns, while in the PDTB only 0.6% of all sentences does not show any relation to other sentences in the text. The LUNA corpus, indeed, contains a lot of disfluencies and semantically empty turns, for example discourse markers, which do not belong to any discourse relation (see Section 4.2.). Besides, a single argument often includes two or more turns when it is expressed discontinuously.
As for the different relation types, the percentage of Explicit relations in the LUNA corpus is much higher than in the PDTB (65.5% vs. 45.75%), while all other types

| Annotated data | |
|---|---|
| N. of dialogs | 60 |
| N. of turns | 3,750 |
| N. of tokens | 24,800 |
| N. of Explicit relations | 1,052 (65.5%) |
| N. of Implicit relations | 487 (30.3%) |
| N. of AltLex relations | 11 (0.7%) |
| N. of EntRel relations | 56 (3.5%) |
| Tot. Annotated relations | 1,606 |
| N. of unique Explicit connectives | 85 |
| N. of unique Implicit connectives | 23 |

Table 1: Statistics about the annotated corpus

are more frequent in the PDTB (39.79% Implicit, 1.55% AltLex and 12.91% EntRel, for details see (PDTB-Group, 2009), pp. 3-4). This might depend on the different annotation procedure adopted for the two corpora: annotators of the PDTB were asked to identify implicit relations only between adjacent sentences, thus leading to significant numbers of EntRels. The LUNA approach, instead, was less strict, and annotators had to identify (implicit and explicit) long-distance relations as well, which are very numerous because of overlaps and interruptions.

In Fig. 2 we report the occurrences of the most frequent connectives, both in implicit and in explicit discourse relations. Note that each connective can be used in different contexts and correspond to different senses, for example "*e*" (*and*) can express a TEMPORAL.*Synchronous* relation, a TEMPORAL.*Asynchronous* relation as well as an EXPANSION.*Conjunction* relation (*semantic* or *speech-act* subtype).

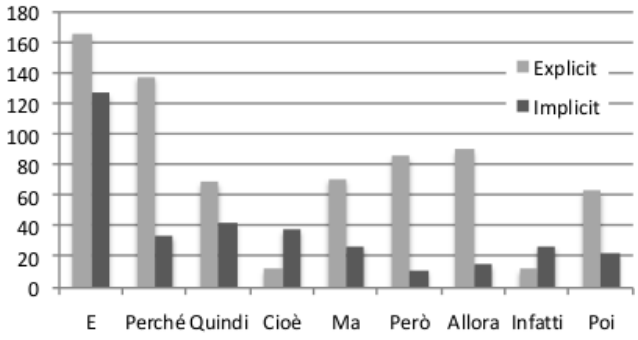


Figure 2: Occurrences of the most frequent connectives

The connectives from left to right are: *and*, *because*, *so*, *that is*, *but*, *instead*, *then* (inferential), *indeed*, *then* (temporal). As expected, the most frequent connective is "*e*" (*and*), followed by "*perché*" (*because*) as explicit connective and "*quindi*" (*so*) as implicit. The ranking is different from the PDTB, where the three most frequent explicit connectives are "*but*", "*and*" and "*because*" (in decreasing order) and the most frequent implicit ones are "*because*", "*and*" and "*specifically*". Another difference between the LUNA annotation and the PDTB is the variability of the connectives: in LUNA 85 unique explicit connectives and 23 implicit connectives were found, while in the PDTB they are respectively 100 and 102. This depends on the different corpus dimensions, but could be explained also in the light of the different nature of the two corpora: in spontaneous dialogs, the speakers seem to use a small set of general connectives to cover a wide range of different relations. This may be due also to the fact that prosody contributes to the identification of the relation. In the newspaper articles belonging to the PDTB, instead, words and expressions are more carefully selected, the vocabulary is richer and connectives tend to be more specific to single discourse relations.

In Fig. 3 the occurrences of the most frequent sense labels are reported separately for explicit and implicit relations. As for the senses, *Repetition* is by far the most frequent relation among implicit connectives (it does not exist as explicit relation because in repetitions a connective is always missing). Among explicit relations, *Temporal.Asynchronous* and *Semantic cause* are the most recurring senses. Anyhow, if we sum the occurrences of *Epistemic cause* and *Speech act cause*, they are as frequent as *Semantic cause*, meaning that our emphasis on pragmatic senses in dialogs is well-founded.

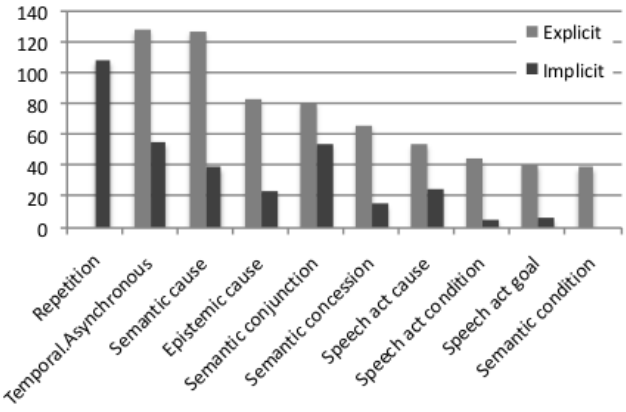


Figure 3: Occurrences of the most frequent sense labels

Even if the PDTB senses are not directly comparable because of the different sense hierarchy, the most frequent sense labels there are *(Semantic) conjunction*, *(Semantic) contrast* and *Reason* (a subtype of *Semantic cause*) for explicit relations, while they are *(Semantic) conjunction*, *Specification* and *Reason* for implicit relations. We believe that the different ranking between the two corpora may depend on the LUNA domain: since the dialogs in the LUNA corpus are typically conversations between a caller and an operator, in which the caller describes a problem with a device and the operator asks questions in order to understand how the problem arose, a lot of turns concern the description of steps carried out to operate a device, which explains the high number of *Temporal.Asynchronous* relations. Also the task-oriented nature of the interaction, with the operator making questions to find out the reason why a problem occurred, can explain the top-ranking of causal relations.

## 6. Conclusions and Future work

In this paper, we have described the annotation of discourse relations in the LUNA corpus. A major goal of our

work was to investigate how the Penn Discourse Treebank (PDTB) annotation framework and its guidelines could be adapted to the annotation of spontaneous conversations in a specific domain and in a new language, namely Italian.
From our initial study, we found that some modifications of the PDTB annotation scheme were required to deal with specific kinds of relations, for example implicit relations between non-adjacent arguments, which are very frequent in spoken language. Other adjustments were introduced in the sense hierarchy in order to take into account the important role of pragmatics in dialogs. A comparison between the sense and connective frequency in the LUNA corpus and in the PDTB confirmed such differences and corroborates our choice to introduce genre-specific adaptations.
An interesting topic that should be investigated to complete the LUNA annotation is the attribution of discourse relations, i.e. whether the relations or arguments are ascribed to the author/speaker of the text or someone else (Wiebe, 2002). The aim of annotating this information is to ascribe beliefs and assertions to the agent(s) making them and has led to interesting results when applied to the PDTB (Prasad et al., 2007). An annotation scheme for attribution based on the PDTB paradigm has already been developed for Italian (Pareti, 2009) and will be applied to create an Italian discourse treebank of newspaper articles (Pareti and Prodanof, 2010). It may be worth studying how to apply it to the complex structure of dialogs.

## 7. Acknowledgements

We would like to thank Livio Robaldo for the useful discussions about the semantics of Italian connectives.
This work was partly funded by the European Commission project LUNA (contract 33549).

## 8. References

C. Bazzanella. 1990. Phatic connectives as Interactional cues in Contemporary Spoken Italian. *Journal of Pragmatics*, 14:629–647.

M. Berretta. 1984. Connettivi testuali in italiano e pianificazione del discorso. In *Linguistica Testuale*, pages 237–253. Bulzoni.

M. Dinarelli, S. Quarteroni, S. Tonelli, A. Moschitti, and G. Riccardi. 2009. Annotating Spoken Dialogs: from Speech Segments to Dialog Acts and Frame Semantics. In *EACL Workshop on Semantic Representation of Spoken Language*, Athens, Greece.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Lucie Mladová, Šárka Zikánová, and Eva Hajičová. 2008. From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco.

U. Oza, R. Prasad, S. Kolachina, D. M. Sharma, and A. Joshi. 2009. The Hindi Discourse Relation Bank. In *Proceedings of the Third Linguistic Annotation Workshop*, Suntec, Singapore.

Silvia Pareti and Irina Prodanof. 2010. Annotating Attribution Relations: Towards an Italian Discourse Treebank. To appear in Proceedings of the Seventh International Conference on Language Resources and Evaluation.

Silvia Pareti. 2009. Towards a discourse resource for Italian: developing an annotation schema for attribution. Master's thesis, Faculty of Letters and Philosophy, University of Pavia, Italy.

R. Prasad, N. Dinesh, A. Lee, A. Joshi, and B. Webber. 2007. Attribution and its annotation in the Penn Discourse TreeBank. *Traitement Automatique des Langues, Special Issue on Computational Approaches to Document and Discourse*, 47(2).

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco.

Silvia Quarteroni, Giuseppe Riccardi, Sebastian Varges, and Arianna Bisazza. 2008. An Open-Domain Dialog Act Taxonomy. Technical Report DISI-08-032, Department of Information Engineering and Computer Science, University of Trento.

PDTB-Group. 2009. The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania. Available online at http://www.seas.upenn.edu/ pdtb/PDTBAPI/pdtb-annotation-manual.pdf.

Deborah Schiffrin. 1987. *Discourse Markers*. Cambridge University Press, Cambridge, U.K.

E. Sweetser. 1990. *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure*. Cambridge University Press.

Sara Tonelli and Giuseppe Riccardi. 2010. Guidelines for Annotating the LUNA corpus with Frame Information. Technical Report DISI-10-017, Department of Information Engineering and Computer Science, University of Trento.

Janyce Wiebe. 2002. Instructions for annotating opinions in newspaper articles. Technical Report TR-02-101, Department of Computer Science, University of Pittsburgh.

Nianwen Xue. 2005. Annotating Discourse Connectives in the Chinese Treebank. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*.

Deniz Zeyrek and Bonnie Webber. 2008. A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. In *Proceedings of IJCNLP*.