LICO: A Lexicon of Italian Connectives

Anna Feltracco
Fondazione Bruno Kessler
University of Pavia, Italy
University of Bergamo, Italy
feltracco@fbk.eu

Elisabetta Jezek
University of Pavia
Pavia, Italy
jezek@unipv.it

Bernardo Magnini
Fondazione Bruno Kessler
Povo-Trento, Italy
magnini@fbk.eu

Manfred Stede
University of Potsdam
Potsdam, Germany
stede@uni-potsdam.de

Abstract

English. This paper presents the first release of LICO, a Lexicon for Italian COnnectives. LICO includes about 170 discourse connectives used in Italian, together with their orthographical variants, part of speech(es), semantic relation(s) (according to the Penn Discourse Treebank relation catalogue), and a number of usage examples.

Italiano. *Questo contributo presenta la prima versione di LICO, un lessico di connettivi per l'italiano. LICO comprende circa 170 connettivi del discorso usati in italiano, di cui abbiamo raccolto varianti ortografiche, le parti del discorso, le relazioni semantiche (ricavate dal catalogo del Penn Discourse Treebank) espresse dal connettivo, e alcuni esempi d'uso.*

1 Introduction

Discourse connectives are explicit lexical markers that are used to express functional relations between parts of the discourse. As an example, the italian word "quando" in the sentence "Quando si preme sul bottone, la porta si apre da sola" (When you press the button, the door opens by itself) expresses a conditional relation between two parts of the sentence (from now on, *arguments*).

Work on discourse connectives in Computational Linguistics was initially part of Rhetorical Structure Theory (Mann and Thompson, 1988), where the focus is on discourse relations, which are at the basis of the notion of textual coherence. In Computational Linguistics, being able to identify connectives is a central task in "shallow discourse parsing", which has become very popular in recent years (e.g., (Lin et al., 2014)) and constituted the shared task of the CONLL conference in 2015 and 2016[1]. Downstream applications that can benefit from shallow discourse structure are, inter alia, sentiment analysis (e.g., (Bhatia et al., 2015) and argumentation mining (e.g., (Peldszus and Stede, 2013)).

Our work on connectives is mainly motivated by the fact that, to the best of our knowledge, still there is no high coverage resource of discourse connectives available for Italian. LICO, the Lexicon for Italian COnnectives, aims at filling this gap, providing a repository of Italian connectives aligned with recent developments in discourse relations (i.e. the last version (3.0) of the Penn Discourse Treebank (PDTB)).

In addition, the LICO lexicon takes advantage from DimLex, a similar repository for German (Scheffler and Stede, 2016; Stede and Umbach, 1998); in fact DimLex served as the main inspiration for creating LICO (see section 4). DimLex is an XML-encoded resource that can be used for NLP; the public version provides information on orthographical variants, syntactic behavior, semantic relations (in terms of PDTB), and usage examples. It is used for automatic discourse parsing, and also for semi-automatic text annotation using the ConAno tool (Stede and Heintze, 2004). Another relevant resource for connectives is LEXCONN, for French, (Roze et al., 2012), which contains about 300 connectives with their syntactic category and coherence relations from Segmented Discourse Representation Theory (Asher and Lascarides, 2003)(and to some extent Rhetorical Structure Theory (Mann and Thompson, 1988)).

LICO is freely distributed under a CC-BY licence.

[1] http://www.cs.brandeis.edu/ clp/conll16st/

2 Discourse Connectives

The definition of discourse connective is controversial both in traditional grammar and in the linguistic literature. Our definition is based on the encyclopedia entry on connectives by Ferrari (2010), included in the reference work for the Italian language recently published by Treccani. In this entry, connectives are defined as "each of the invariable forms [...], that introduce relations that structure "logically" the meanings of the sentence and of the text"[2]. The definition provided in Ferrari (2010) is restrictive, as it does not include variable forms, i.e. those forms which are subject to morphological modifications, such as *ne consegue/conseguiva che* 'it follows/followed/ that', nor does it include pragmatic uses of connectives (also known as *discourse markers*) such as causal *perché* 'why' in "Che ore sono? *Perché* ho dimenticato l'orologio" ('what time is it? Because I forgot my watch'). On the other end, it assumes that logical relations marked by connectives hold between events or assertions, and therefore includes as arguments for the relation nominal expressions such as "dopo il pressante invito ..." 'after the pressing invitation ...', i.e. expressions that contain an event nominal, - although the event is, in this case, referred to instead of predicated.

In our work, we partly drop the invariability criteria; we do not include forms which exhibit morphological inflection or conjugation, but we do include connectives which show a certain degree of lexical variability that is, multi-word expressions which are not totally rigid from a lexical point of view (*ad esempio/per esempio* 'for example'; see section 3).

3 The Structure of the Lexicon

Each entry in the LICO lexicon corresponds to a connective (including its variants). Currently, for each entry LICO specifies:

- whether the connective (or its variants) is composed by a single token ("part = single", e.g. *perché*) or by more than one token ("part = phrasal" e.g. *di conseguenza*);
- whether the connective is composed by correlating part ("orth = discont") or not ("orth = cont") and the specification of the two correlating parts, e.g. "orth = discont": *da una parte* ("part = phrasal"), *dall'altra* ("part = phrasal"); "orth = cont": *perché* ("part = single");
- possible orthographic variants: e.g. *ciò nonostante* ("part = phrasal") and *ciononostante* ("part = single");
- possible lexical variants: e.g *dopo di ché* and *dopo di ciò*. Notice that in some cases this lexical variants determine a different syntactic environment, such as *in modo da* and *in modo che*, the first being followed by infinitive form, the following by a subjunctive form;
- pos category: adverbs, preposition subordinating or coordinating conjunctions;
- the semantic relation(s) that the connective indicates, according to the PDTB 3.0 schema (see section 3.1);
- examples of the connectives for each semantic relation;
- possible alignments with lexicon of connectives in other languages.

Table 1 shows the entry for *quando*, which presents more than one semantic relation, and the entry for *ciononostante, ciò nonostante, nonostante ciò*, as example of a connective with orthografic variants in LICO.

3.1 Semantic relations

For the annotation of the semantic relation we used the PDTB 3.0 schema of relations (Webber et al., 2016; Rehbein et al., 2016) as proposed in the DimLex resource (Scheffler and Stede, 2016), which is our main reference resource.

The schema is a most recent version of PDTB 2.0 (Prasad et al., 2008; Prasad et al., 2007) and includes semantic relations structured in a hierarchy composed by three levels. In the first level, the *class level*, the relations are grouped in four major classes: *TEMPORAL*, *CONTINGENCY*, *COMPARISON* and *EXPANSION*. The second level, the *type level*, specifies further the semantics of the class level. For example, the *TEMPORAL: Synchronous* tag is used for connectives that indicate that the two arguments are simultaneous, while the

[2] "Il termine *connettivo* indica in linguistica ciascuna delle forme invariabili [...], che indicano relazioni che strutturano 'logicamente' i significati della frase e del testo".

▷ entry-id	146
▷ orth	cont
▷ part	single
	quando
▷ POS	subordinating
▷ sem relation	TEMPORAL: Synchronous ex.: Quando lasciò l'appartamento, arrivò la chiamata rel. to German id: 5
▷ sem relation	CONTINGENCY:Condition ex.: Quando si preme sul bottone, la porta si apre da sola. ex.: Quando me lo chiedi, lo lascerò stare. rel. to German id: 116
▷ entry-id	30
▷ orth	cont
▷ part	single
▷ variant	orthographic
	ciononostante
▷ orth	cont
▷ part	phrasal
▷ variant	orthographic
	ciò nonostante
▷ orth	cont
▷ part	phrasal
▷ variant	orthographic
	nonostante ciò
▷ POS	coordinating
▷ sem relation	COMPARISON:Concession:Arg2-as-denier ex.: La procura ha ordinato la restituzione dell'esemplare confiscato. Ciononostante l'istruttoria prosegue. rel. to German id: 74

Table 1: The connectives *quando* and *ciononostante, ciò nonostante, nonostante ciò* in LICO.

TEMPORAL: Asynchronous tag is used for connectives that indicate a before-after relation between the arguments. The third level (*subtype level*)[3] varies according to the role of the two arguments involved in the relation. For example, *CONTINGENCY:Cause:Reason* is used if the argument introduced by the connective -Arg2- is the reason for the situation in the other argument -Arg1- (e.g. I stayed at home, because it was raining), while *CONTINGENCY:Cause:Results* is used if Arg2 represents the result/effect of Arg1 (e.g. It was raining, therefore I stayed at home). Not every *type* has a further *subtype*.
In the LICO structure, each connective is assigned with one or more three-level tags.

4 The Current Resource

In this Section, we present the current resource and its construction. In particular, we focus on describing how the list of entries has been identified so far and how we proceeded to acquire the semantic information for each entry.

List of connectives. Currently, LICO is composed by 173 entries, each one corresponding to a connective and its orthographical or lexical variants. In order to compile this list we used a number of grammatical and lexical resources for Italian and for other languages.

First, we retrieved the list of connectives mentioned by Ferrari (2010) in the Enciclopedia Treccani for the entry *connettivi*[4] for a total of 33 connectives. Then, we retrieved the list of connectives tagged as *congiunzione testuale* in Sabatini Coletti 2006 (Sabatini-Coletti, 2005) discarding the ones of literary use, for a total of 70 entries. Finally, we benefited from the DimLex resource for German, as we enriched our list by identifying the equivalent Italian terms of the German connectives[5]. This process was facilitated by the presence of examples in the German resource in which the connective is displayed in context: only the Italian candidates that maintain the sense of the German connectives were added to LICO. We keep trace of this "German-Italian" links and we will use this information to enrich also the characteristic of the entry in LICO (e.g. *aber* → *ma*). A total of 127 entries were collected with this method. Figure 1 shows the overlap between the three resources and Table 2 shows a sample of the connectives in LICO and the respective sources.

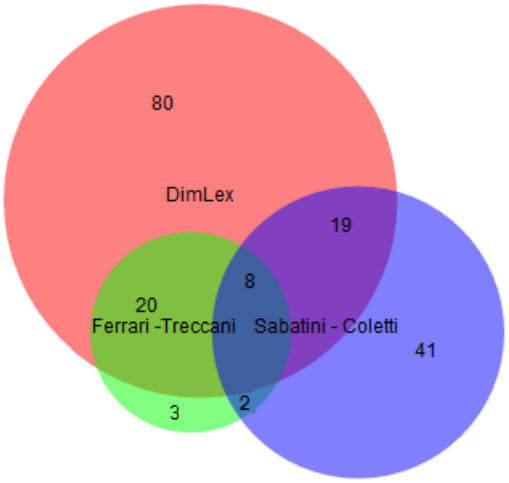

Figure 1: Overlap between the resources.

Semantic relations in LICO.
In LICO connectives are tagged with the semantic relations that the connective can indicate in a text, selecting the most appropriate ones in the PDTB 3.0 schema. In this process we took advantage from the information which was already

[3]The names of the levels are taken form Prasad et al. (2007).

[4]http://www.treccani.it/enciclopedia/connettivi_(Enciclopedia-dell'Italiano)/, last access July 21st 2016.

[5]https://github.com/discourse-lab/dimlex

LICO Entries	Resources		
	Ferrari Treccani	**Sabatini Coletti**	**DimLex (equivalent)**
dopo	dopo	dopo	dopo
dopo di che dopodiché		dopodiché	dopo di che
dopotutto		dopotutto	
dunque	dunque	dunque	dunque
e	e		e
ebbene		ebbene	
eccetto			eccetto
eppure		eppure	eppure

Table 2: Sample of connectives in different resources.

present in the resources we used for building the list. In fact, the DimLex resource provides this information for the German connectives, and both the Italian resources previously mentioned provide useful information about the semantic relation triggered by the connective.[6] A total of 23 different PTDB relations have been used to describe LICO entries. In order to validate the tagging of semantic relations, we conducted a research by observing examples of the use of the connectives in corpora, i.e. we wanted to verify whether the relation that a connective introduces in a portion of text is one of the relations already tagged for that same connective in the first step. In particular, we searched for 20 connectives in the ItWac corpus (Baroni et al., 2009) and we retrieved occurrences with 400 characters on both sides of the connective. We limited our observation to 5 retrieved segments of text in which the connective is actually playing such a role. We finally tagged each connective in each portion of text with the semantic relation it indicates.

To further confirm the corpus-driven evidences for the semantic relations, we asked two annotators (one being an expert annotator, the other not) to perform the same tagging task. We then calculated the interannotator agreement between the two annotators adopting the Dice's coefficient (Rijsbergen, 1997)[7] for three configurations, one for each level of the relation schema: *class agreement*, *type agreement*, *subtype agreement*. We considered that there was agreement if both annotators identify exactly the same *class*, *type*, *subtype* respectively. The Dice values result in 0,78 for *class agreement* and 0,71 for both *type agreement* and *subtype agreement*.

Observing cases of disagreement, we can make the following preliminary considerations. The main cases of disagreement regard the COMPARISON:Contrast relation (on one hand) and the COMPARISON:Concession and EXPANSION:Substitution relations (on the other hand). These relations in fact appear to be the ones that connect arguments that are in contrast. As an example, the connective *anziché* 'rather than' in Example (1) has been annotated as COMPARISON:Contrast by annotator1 and as EXPANSION:Substitution:Arg1-as-subst by annotator2: the first enlightens the contrast between "emissione attraverso il Tesoro" and "usare il tradizionale sistema", the second emphasises that Arg2 represents the alternative to the Arg1.

(1) [..] chiedeva l' emissione di dollari in banconote statunitensi attraverso il Tesoro *anziché* usando il tradizionale sistema della Federal Reserve.

Another interesting case concerns the disagreement between the relations TEMPORAL:Asynchronous:precedence (in which Arg2 follows Arg1) and CONTINGENCY:Cause:Result (in which Arg2 is the results of Arg1), being the two strictly connected (i.e. in a cause-effect relation, the effect follows the cause). As an example, in (2) one annotator marks the connective as indicator of the temporal sequence of Arg1 and Arg2, while the other prefers to mark it as an indicator of the cause-effect relation.

(2) [..] Il bello è che i tipi hanno pure accennato a prendersela con me, *al che* io gli ho abbaiato contro una sequela di insulti [..]

In general, the relations that were initially as-

[6] In particular, in the online version of Sabatini Coletti (http://dizionari.corriere.it/dizionario italiano/D/dizionario .shtml, last access July 21st 2016) the semantic relations the connectives can trigger are described in the definition of the connective itself, e.g. "*quindi*, cong. testuale: Con valore deduttivo-conclusivo, perciò, di conseguenza, per questo motivo, dunque". Ferrari (2010) in the Enciclopedia Treccani proposes a non hierarchical classification which includes the following relations: "temporal relation" "causal relation", "consequence relation", "condition relation", "opposition relations".

[7] Dice's coefficient measures how similar two sets are by dividing the number of shared elements of the two sets by the total number of elements they are composed by. This produces a value from 1, if both sets share all elements, to 0, if they have no element in common.

signed to these connectives were confirmed by the corpus-based exercise (i.e. at least one annotator assigns the tag in at least one portions of text); viceversa, in some cases one of the two annotators assigned a relation that was not initially identified.[8]

5 Conclusion and Further work

In this paper we have presented LICO, a new resource for the Italian language describing lexical properties of discourse connectives. While LICO fills a gap with respect to similar resources existing for other languages, it is still under construction under several aspects. Our short term plans include the completion of the lexical entries with corpus derived examples and the observation of the connectives in Italian corpora, in order to acquire more information about the semantic relations that each connective can indicate and thus extend the annotation of the semantic relations in LICO.

Acknowledgment

We acknowledge Denise Pangrazzi for her contribution to identify the Italian equivalents of the German connectives.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, September. Association for Computational Linguistics.

Angela Ferrari. 2010. Connettivi. In *Enciclopedia dell'Italiano*. diretta da Raffaele Simone, con la collaborazione di Gaetano Berruto e Paolo D'Achille, Roma, Istituto della Enciclopedia Italiana.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184.

[8]For this moment, the "new" relations are not included in LICO.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The Penn Discourse Treebank 2.0 Annotation Manual.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating Discourse Relations in Spoken Language: A Comparison of the PDTB and CCR Frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May.

CJ van Rijsbergen. 1997. Information retrieval. 1979.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. LEXCONN: a French lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique.*, (10).

Il Sabatini-Coletti. 2005. Dizionario della lingua italiana 2006, con CD-ROM. *Milano, Rizzoli Larousse*.

Tatjana Scheffler and Manfred Stede. 2016. Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May.

Manfred Stede and Silvan Heintze. 2004. Machine-assisted rhetorical structure annotation. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 425–431, Geneva.

Manfred Stede and Carla Umbach. 1998. Dimlex: A lexicon of discourse markers for text generation and understanding. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1238–1242. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A Discourse-Annotated Corpus of Conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31. Association for Computational Linguistics.