Theoretical Implications of Automated Discourse Parsing in Student Writing

Arianna Bienati*
Università degli studi di Modena e Reggio Emilia

Mariachiara Pascucci**
Università di Pisa

Jennifer-Carmen Frey†
Eurac Research

Alessio Palmero Aprosio‡
Università di Trento

This article presents a study on annotating explicit discourse relations in Italian student essays, comparing human annotations with outputs from generative large language models and examining their alignment with theoretical models of textuality. We review prior work on automatic discourse relation annotation in Italian, highlighting limitations in language coverage, especially in out-of-domain scenarios and how these have been addressed. Our experiments explore the use of generative models to mitigate the scarcity of domain-specific training data, while assessing their ability to reflect the intended theoretical framework. We evaluate two generative models in detecting connectives and classifying their senses, comparing results to human annotation. For our evaluation sample, we use a string-matching algorithm combined with a rule-based approach to pre-annotate essays with possible connective forms and their senses, based on their presence in the Lexicon of Italian Connectives (LICO). These annotations were manually corrected by two expert annotators, resulting in a publicly available evaluation sample. The study raises significant theoretical questions about the definition of connectives, its relationship to text segmentation and the challenges both human and machines face when annotating discourse relations. Our findings show how computational approaches can shed light on linguistic theories and, vice versa, how linguistic theories can guide the application of computational resources.

1. Introduction

Textual competences are essential for writing high quality texts. This becomes particularly evident as writing tasks become more demanding and texts increase in length — such as in final school examinations or at various stages of university coursework (e.g., papers, reports, and theses). Despite their centrality, students' textual competences have been found lacking in various empirical investigations, including research focused on the Italian language, such as Cisotto and Novello (2012). These pedagogical studies, however, seldom provide quantitative evidence to support their claims. In contrast,

* Dipartimento di Educazione e Scienze Umane - Viale Timavo 93, 42121 Reggio Emilia, Italy. E-mail: arianna.bienati@unimore.it
** Dipartimento di Filologia, Letteratura e Linguistica - piazza Evangelista Torricelli 2, 56126 Pisa, Italy. E-mail: mariachiara.pascucci@phd.unipi.it
† Institute for Applied Linguistics - Viale Druso 1, 39100 Bolzano, Italy. E-mail: jennifer-carmen.frey@eurac.edu
‡ Dipartimento di Psicologia e Scienze Cognitive, Università di Trento - Corso Bettini 84, 38068 Rovereto (TN), Italy. E-mail: a.palmeroaprosio@unitn.it

research that does offer data-driven insights into student writing (Barbagli et al. 2015; Sprugnoli et al. 2018; Ruele and Zuin 2020) — often based on large corpora of essays annotated through both computational and manual methods — tends not to concentrate on textual competences. Rather, such work typically addresses other aspects of linguistic analysis, such as orthography, vocabulary, syntax, and traits of the Italian 'neostandard' variety.

In the context of a research project aimed at describing students' textual competences at the end of mandatory schooling[1], we attempted to fill this gap by investigating how students typically construct texts in terms of explicit discourse relations, which are an important component of text construction alongside other components such as anaphoric relations, topic-comment progression, and the alternation of different voices and perspectives. For this purpose, a corpus for evidence-based analysis of textual competences of students should be enriched with annotations of explicit discourse relations using strategies of discourse parsing.

Discourse parsing is an NLP task that aims to automatically annotate the discourse structure of a text according to a specific 'discourse grammar'. There are various discourse theories and grammars with which a text can be parsed. Among the most common and widely used frameworks, there are Rhetorical Structure Theory (RST, Mann and Thompson (1988)), whereby the text is first segmented into elementary discourse units and then structured into a tree structure, and D-LTAG (Webber 2004), i.e. Lexicalised Tree Adjoining Grammar applied to the discourse level. The theoretical work produced in the D-LTAG framework gave the foundations to the Penn Discourse Treebank (PDTB, Miltsakaki et al. (2004)), a manually annotated, comprehensive resource for the English language. Differently from RST, the PDTB builds on a flat discourse structure between clauses or sentences, whose nodes of conjunction are lexically (i.e. explicitly) realized or implicitly realized predicates. Predicates can be either connectives, i.e. coordinating or subordinating conjunctions and specific types of adverbials that link two arguments on a discourse level (e.g., *and*, *because*, *however*), alternative lexicalizations, such as the phrasing *the reason is*, or they can be left implicit. Predicates operate on linguistically specified arguments (usually referred to as *arg1* and *arg2*), which in turn should be identified and labeled in order to build the discourse structure.

For the Italian language, research on discourse parsing has centered around the PDTB framework and its shallow discourse structure, identifying discourse relations, both signalled by connectives and alternative lexicalizations or left implicit, and subsequently categorizing them into different sense categories according to the PDTB taxonomy. Approaches to discourse parsing so far were heavily based on the manual annotation of resources that were then used to train supervised models on the task (Pareti and Prodanof 2010; Tonelli et al. 2010; Feltracco, Magnini, and Jezek 2017). Alternative approaches also considered lexical resources such as dictionaries of connectives to approximate similar results in low-resource scenarios (Feltracco et al. 2016). Approaches prompting generative Large Language Models (LLMs) have barely been considered yet, although they could offer a viable solution for languages and domains in which no comprehensive training data is available.

In our study, we make use of both lexical resources and generative models to annotate explicit discourse relations for the following reasons. Firstly, our data cover a

1 "ITACA – Coerenza nell'ITAliano Accademico", funded by Provincia Autonoma di Bolzano (Ripartizione "Diritto allo studio, università e ricerca scientifica". Legge provinciale 13 dicembre 2006, n.14 "Ricerca e innovazione"). For a description see Zanasi et al. (2024). The corpus is accessible at `https://www.porta.eurac.edu/lci/itaca/`.

peculiar domain which is different from the domains for which annotated resources are available (usually news, legal documents, Wikipedia pages, and social network posts). Secondly, LLMs and few shot prompting techniques provide the necessary flexibility to handle custom annotation schemas. The ITACA project entailed the annotation of multiple discourse phenomena under the unifying framework of the Basel Model of Textuality (Ferrari 2014; Ferrari, Lala, and Zampese 2021). While the sense categorization of connectives in the Basel Model departs only minimally from the PDTB senses — which could therefore be retained — the criteria identifying explicit connectives depart more evidently. To maintain coherence with other annotations, we thus decided to adopt the Model's connective definition and to test whether expert annotators would be able to perform reliable annotations using this definition as the first step of the study. Only after this goal was met, an automatic annotation was considered. We evaluate the performance of LLMs in identifying and categorizing discourse connectives against the backdrop of Ferrari's theoretical model of connectives in Italian student essays collected in Italian upper secondary schools from the autonomous province of Bolzano. We test using it for two different prompt versions (with and without connective and sense category definitions) in a few shot scenario on both GPT-4o and Llama 3.3 70B, and compare resulting annotations with the manual annotations from the human annotators. Finally, we investigate frequent errors in both human and automatic annotation and relate them to each other. In total, our study addresses the following research questions:

1. What is the level of human agreement in detecting and classifying explicit discourse relations in students' essays basing on Ferrari's (2021) definition?
2. Which are the most challenging relations and connective words?
3. How do generative transformer models perform on the same tasks?
   - Can providing connective and sense category definitions improve model performance?
   - Do the models' mistakes align with human disagreements?

The article is structured as follows: Section 2 gives an overview of prior research on shallow discourse parsing in Italian, while Section 3 introduces our data and manual annotation task. Experiments with LLMs are presented in Section 4 and results of both the manual annotation task and the LLMs experiments are reported in Section 5. We offer in-depth discussion of disagreements and misclassifications in Section 6, followed by a summary with limitations and some ideas for future research (Section 7). Information about the release of the data (Section 8) are provided at the end of the paper.

2. Shallow discourse parsing for Italian

Automatic approaches to shallow discourse parsing have usually relied on supervised training (Pitler and Nenkova 2009), which requires substantial amounts of manually annotated data that sufficiently represent all possible types of relations and, ideally, span various language domains to ensure model transferability. While the English PDTB 3.0 (Prasad, Rashmi et al. 2019) is the most comprehensive resource available for discourse annotation, covering implicit and explicit discourse relations of all types with satisfactory inter-annotator agreement, other languages are less well-represented. For the Italian language for example, manually annotated resources cover mainly subsections of the framework, usually focusing on specific discourse relations. The Italian Syntactic-Semantic Treebank (Montemagni et al. 2000), for example, has been enriched with anno-

tations for attribution relations (Pareti and Prodanof 2010), making improvements to the PDTB attribution relation annotation scheme. Contrastive and concessive relations, on the other hand, are the subject of the Contrast-ITA bank (Feltracco, Magnini, and Jezek 2017), a resource consisting of 169 newspaper articles from the I-CAB corpus (Magnini et al. 2006), in which annotations for implicit, explicit and alternative lexicalizations of contrastive and concessive relations can be found. The most comprehensive resource at the level of types of annotated discursive relations is the LUNA corpus (Tonelli et al. 2010), which contains annotations for implicit, explicit and alternatively lexicalized relations for all four possible macro-senses of the PDTB (temporal, contingency, comparison and expansion). The corpus, however, contains only dialogues; it is thus specialised in the type of discourse structure typical of the oral mode.

Manually annotating discourse relations is not only laborious but also requires clear and unambiguous definitions of what constitutes the elements to be annotated (e.g. what counts as a connective and what counts as a possible argument). While theoretical models might be vague in some points, derived annotation manuals must clarify ambiguities as much as possible, often leading to documents of extensive lengths and various revisions of both manuals and annotated resources. The English PDTB, for example, already reached its third revision and its manual annotation process (from its second version onwards) has been guided by two annotation manuals (Prasad et al. 2007; Webber et al. 2019). Even within the PDTB annotation style alone, there are different ways of segmenting and detecting connectives. While the resources annotated in this style overlap in terms of the syntactic categories and most of the functional criteria that define an instance (be it word or multi-word expression) as connective, there are substantial differences on the segmentation practices declared in the various annotation manuals. For instance, in PDTB 2.0 (Prasad et al. 2007), coordinating conjunctions that coordinate verb phrases (VP coordinations) are not considered connectives (e.g., in the sentence "It employs 2,700 people and has annual revenue of about $370 million.", *and* is not a connective). However, adverbial forms that appear in VP coordinations do (e.g., "It acquired Thomas Edison's microphone patent and *then* immediately sued the Bell Co."). The LUNA corpus, even if it follows the PDTB 2.0 Annotation Manual, considers these coordinating conjunctions connectives (e.g., "Chiede di effettuare la connessione a internet *o* inserire il supporto rimovibile", *It requests to connect to the internet or insert removable media*). Similarly, in the transition between PDTB 2.0 and 3.0 (Webber et al. 2019), coordinating conjunctions in VP coordinations are considered connective words *per se* (e.g., "It employs 2,700 people *and* has annual revenue of about $370 million.": now *and* is a connective). Another important difference regards the annotation of *to*-infinitive clauses: in the PDTB 2.0 they are not considered arguments of a discourse relation, whereas in PDTB 3.0 they count as possible arguments (e.g., "The Galileo project started in 1977, and a number of project veterans were on hand *to watch the launch*.").

Once sufficient inter-annotator agreement is reached, resources can be used for training supervised models. Since the release of the second version of PDTB, there have been various studies that have trained automatic systems for discourse parsing, partly focusing on specific sub-tasks, such as connective detection or relation classification for either implicit, explicit or all types of discourse relations. For English, already simple approaches (e.g., max entropy classifiers with syntactic or lexical features, Pitler and Nenkova (2009)) manage to achieve ceiling performances. The 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023) (Braud et al. 2023) is the first instance where an Italian dataset, the LUNA corpus, was used as training data to perform connective detection and relation classification. The best models for Italian showed performances between $\mathrm{F1} = 65.85$ and 71.60 for connective detection

(cf. Metheniti et al. (2023)) and an F1 between 65.00 and 72.48 (depending on whether implicit relations were considered) for a 4-way relation classification, i.e., using only the four course-grained main categories *Comparison*, *Contingency*, *Expansion*, and *Temporal* (cf. Liu, Fan, and Strube (2023)).

Studies have shown that, even for English, performance drops drastically across domains (Scholman et al. 2021) and when averaging over connectives types instead of instances (Johannsen and Søgaard 2013). Relatively poor performances can also be seen when only small annotated datasets are available, as it is the case for all participating models of the DISRPT shared task when applied to the Italian LUNA corpus, representing atypical training data from spoken language. Thus, in order to by-pass the insufficient performance and coverage of domains and languages caused by the lack of comprehensive training data, various approaches have been considered for Italian so far, which we will detail below.

2.1 Lexical resources

One way to aid manual annotations of extensive sets of data for the training of supervised models is to rely on connective lexicons. The TextLink working group has coordinated a series of projects aimed at constructing lexicons of connectives, specific to each language, but related to each other (Stede, Scheffler, and Mendes 2019). These lexicons all refer to the PDTB annotation style and serve to assist explicit discourse connective identification and sense assignment. This method is applicable straightforwardly with string-matching approaches to any kind of texts, across different domains. However, the precision of the annotation can be severely degraded (Bienati et al. 2023). For Italian, LICO (Feltracco et al. 2016) represents the most complete and up-to-date resource for lists of connectives. It contains 173 connectives. For each one, information is given on the syntactic category, the senses that the connective can express, some examples for each sense, the corresponding ones in the German lexicon (DiMLex, Stede and Umbach (1998)) and, if present, it also gives information on the orthographic variants of that connective. LICO has been composed from: a list of connectives taken from the entry about connectives in the Enciclopedia Treccani (Ferrari 2010); other connectives automatically extracted from the Sabatini Coletti dictionary (all lemmas labeled as textual conjunctions were extracted) (Sabatini and Coletti 2005); remaining connectives translated from the German lexicon DiMLex. The LICO also explicitly states a theoretical basis for its definition of connective, which is the one proposed in Ferrari (2010) (see also refinements in Ferrari (2014) and Ferrari (2021)).

2.2 Fine-tuning multilingual models and data augmentation techniques

The first venue in which an Italian dataset, the LUNA corpus, was used as training data to carry out connective identification and sense classification is the 2023 DISRPT shared task (Braud et al. 2023). Both models performing best on Italian data made use of multilingual language models to overcome issues of missing or small training data in the connective detection and relation classification tasks. The DisCut model performing best in the connective detection task used the multilingual XLM-RoBERTa-large (Metheniti et al. 2023) and fine-tuned it on Italian, freezing some of the lower layers known for encoding morpho-syntactic information. Instead, for the 4-way relation classification, the best performing HITS model (Liu, Fan, and Strube 2023) used XLM-RoBERTa-large and fine-tuned it on aggregated training data in Italian, Portuguese, Turkish and Thai.

Even better results in the 4-way relation classification (F1 = 72.72, including implicit and explicit relations) were obtained by Bourgonje and Demberg (2024) using the same model as Liu, Fan, and Strube (2023) but augmenting the Italian dataset with data obtained from translating the English PDTB into Italian and adding it to the training data, as well as adding simple domain adaptation (Daumé III 2007). They also report results for the more complex 11-way classification task, i.e., using both level 1 and level 2 of sense categories in the PDTB, with F1 being 57.13 when considering both explicit and implicit relations.

2.3 Connective Detection with LLMs

The application of Large Language Models (LLMs) to the task of extracting discourse connectives, particularly in Italian, remains an underexplored area in current research. While LLMs have demonstrated proficiency in various natural language processing tasks, their effectiveness in identifying discourse connectives in corpora of student essays has not been extensively investigated. Notably, the DiscoFLAN system employed a Flan-T5 model for connective identification as part of the DISRPT 2023 Shared Task, marking one of the few instances where an LLM was explicitly used for this purpose (Anuranjana 2023; Braud et al. 2023).

Finally, some studies have explored the use of transformers for discourse connective detection in multilingual contexts. For instance, research utilizing BERT-based models has shown promising results in identifying discourse connectives across languages such as English, Turkish, and Mandarin Chinese (Chapados Muermans and Kosseim 2022).

3. Data

In the following, we describe the data used in this study. We introduce the corpus of student essays used and the evaluation set drawn from it as well as the annotation task performed by highly trained human annotators.

3.1 The ITACA Corpus

The ITACA corpus is a collection of argumentative essays written by Italian upper secondary school students of 12th grade. The corpus was collected for the project ITACA - Coerenza nell'ITAliano Accademico (Zanasi et al. 2024, in particular §3.2) that was conducted from 2021 to 2024 and aimed to analyse students acquisition of academic language with a focus on aspects of coherence in their written productions.

All texts were collected during the school year 2021/2022. Students from 12th grade from 13 Italian upper secondary schools (in total 41 classes) in the autonomous province of Bolzano were asked to write an argumentative essay of at least 600 words, given a predefined writing task. Students were asked to write a letter to the Minister of Education, taking position and elaborating on the topic of remote learning (Italian *didattica a distanza*) based on their personal experience during the pandemic and on information they obtained from input texts that were provided to them with the task assignment. The writing task was timed, giving students an upper time limit of 100 minutes to finish the task. Next to providing typed essays for the writing task, students also filled an additional socio-linguistic questionnaire that recorded questions regarding their basic socio-demographic information and their language background and reading and writing habits.

The corpus consists of 635 essays, amounting to a total of 424,693 tokens, with an average text length of 669 tokens per essay. All essays and metadata were collected digitally, using the online tool Survey Monkey[2]. All data has been fully anonymized, ensuring that neither texts nor metadata entries can be traced back to single individuals. After anoymisation, the corpus underwent basic natural language processing (NLP), including automatic tokenization, lemmatization, part-of-speech tagging and dependency parsing using the Italian open-source NLP suite Tint (Palmero Aprosio 2021). At the same time, additional manual annotation was performed using the annotation tool INCEpTION[3] (Klie et al. 2018) for a subset of 388 texts, with the intention to offer more detailed descriptions of text structure and of specific linguistic features related to coherence and cohesion in the texts.

In addition to the corpus, the ITACA project also produced a validated rating scale for assessing coherence in student essays, along with coherence ratings provided by two independent raters who applied this scale to the texts in the corpus.

3.2 Annotation task

For this study, we randomly selected 40 texts from the ITACA corpus to be annotated with the presence, position, and sense categories of connectives. The first five texts served as a pilot to test and refine the annotation guidelines and were therefore excluded from the main analysis. Since both the original manual annotation and the current study involved similar—but not identical—tasks related to connectives, we excluded any texts that were part of the previously annotated set of 388 texts. This was done to avoid bias from annotators who participated in both annotation efforts. Consequently, we selected essays solely from the 247 texts that had not yet been manually annotated. The final evaluation set was then annotated by two annotators within a new INCEpTION annotation project.

As a theoretical basis for the detection of connectives we followed the definition in Ferrari (2021), within the underlying theoretical model described also in Ferrari (2014) and Ferrari, Lala, and Zampese (2021). This definition is quite restrictive, as it limits the category to "ciascuna delle forme morfologicamente invariabili che offrono istruzioni su come legare gli eventi evocati dal testo o gli atti linguistici di composizione testuale attraverso relazioni logico-argomentative"[4] (Ferrari 2021, 145-146). The definition is thus composed of a formal criterion (morphological invariability) and a functional or semantic criterion, which limits connectives to those words that provide procedural information and operate on the logical-argumentative level. A number of restrictions also follow from this definition. Connectives cannot be relative pronouns or complementizers, as they are elements that are invariable, but do not convey logical-argumentative relations. Moreover, prepositions that are linked to first-order semantic entities (such as people, places, things, which exist in time) are not to be considered connectives. Lastly, expressions that, although associated with a logical-argumentative

2 https://www.surveymonkey.com/

3 https://inception-project.github.io/

4 "each of the morphologically invariable forms that provide instructions on how to connect the events evoked by the text or the linguistic acts of textual composition through logical-argumentative relations". The criterion of morphological invariability is reasonable, if only for its ability to distinguish between purely lexical and lexicogrammatical cohesive devices. From a purely practical perspective, having a formal criterion helps with annotation, because it ensures that membership of the class can be easily verified. However we acknowledge that adding this criterion to a fully functional definition is only one of many current conceptions in defining connectives.

relation, are morphologically variable, are not to be considered connectives because they violate the formal criterion expressed above (e.g., *per questo motivo* (English *for this reason*) can be inflected in *per questi motivi* (English *for these reasons*) and lexical insertions are possible, as in *per tutti questi motivi*, English *for all these reasons*).

To speed up the annotation and reduce error, specific tokens were pre-annotated as connectives, using an extended version of the Lexicon of Italian Connectives (LICO, Feltracco et al. (2016)), which stated to follow the same theoretical definition proposed by Ferrari (2010). For this pre-annotation, we used a combination of string matching and rule-based filtering based on part-of-speech tags and dependency relations. Pre-annotated texts and scripts are available on Github (see Section 8 for further details). With regard to annotating sense categories, we followed the PDTB-3 Annotation Manual (Webber et al. 2019), originally developed for the annotation of the English PDTB.

Thus, annotation process consisted of two main tasks:

1. determining whether pre-annotated tokens constitute a connective – this first task also involved manually adding annotations of connectives if not captured in the pre-annotation;
2. identifying the appropriate sense category from scratch using the PDTB annotation manual – this task was performed without relying on any pre-annotation guidance for annotators.

Comprehensive annotation guidelines are released together with the dataset (see Section 8). After completing the annotation procedure, any disagreement cases were discussed in a reconciliation session, ultimately establishing a curated version of the documents. Results of the manual annotation task are illustrated in Section 5 together with the results of the experiments with LLMs.

4. Experiments

In our experiments, we explore a variety of approaches and techniques, to simulate manual annotation using a generative model. Given the complexity and rarity of the annotation categories, we adopt a few-shot learning strategy, providing the model with a small set of examples (up to five) for each category involved in the annotation process.

The dataset is divided into two subsets: training and testing. Out of the 35 documents in the dataset, 25 are used for training and the remaining 10 for evaluation. The training set is specifically employed to select examples for the few-shot prompts, while the test set serves to assess the model's performance.

Train/test splits are created using Iterative Stratification (Sechidis, Tsoumakas, and Vlahavas 2011; Szymański and Kajdanowicz 2017), which ensures a balanced distribution of concepts across the partitions. In our case, a "concept" refers to a word-annotation pair, with "not a connective" being treated as a valid annotation.

4.1 Long and short prompt

We experiment with two prompting strategies for the generative model. In the first approach, we include detailed definitions of what constitutes a connective, along with descriptions of the various categories used for classification. The second approach takes a more minimalistic route, relying solely on a set of example sentences drawn from the few-shot setup, without any additional explanation.

The two prompting strategies were designed not merely to contrast verbosity, but to investigate the impact of theory-rich vs. example-driven instruction on model performance. The long prompt includes explicit definitions and category descriptions derived from linguistic theory (Ferrari 2021; Webber et al. 2019), aiming to simulate the guidance typically provided in annotation manuals. The short prompt, on the other hand, relies solely on annotated examples, reflecting a more empirical, data-centered approach. This contrast allows us to explore whether theoretical framing enhances the model's understanding of the task, or whether exposure to labeled examples alone is sufficient.

The full prompts are provided in Appendix A.

4.2 One-by-one and Full-text

We define two different annotation strategies for connectives: (i) starting from a specific expression (one-by-one), and (ii) annotating an entire document (full-text).

In the first strategy, a target expression from the LICO dataset (e.g., *ma*) is selected. Using the training data, we then build a few-shot prompt by:

- extracting all sentences in the training set that contain the chosen expression;
- identifying all annotation labels assigned to that expression;
- for each pair expression/label, randomly selecting up to five sentences containing the expression and annotated with the label.

To make this process more efficient, we only consider expressions that have been annotated with at least two different labels, including "not a connective" (e.g., *ma* can be labeled as Comparison:Contrast or Comparison:Concession or *prima di* can be labeled as Temporal:Asynchronous or "not a connective"). This ensures that the model is not tasked with trivial classifications where all few-shot examples would belong to the same category.

When choosing sentences from the training set for the few-shot approach, we only include those with annotation agreement among annotators. This prevents the generative model from encountering ambiguous data that could make the task harder.

In the second strategy, a text from the test set is selected, and the list of potential connectives identified using the heuristics described in Section 3.2. If a candidate connective is associated with only one category in the training set, it is excluded, as its classification would be considered straightforward. For the remaining candidates, a set of examples is retrieved from the training set — following the approach used in the first strategy — and incorporated into the prompt.

4.3 Choice of the LLM

We evaluated both strategies using two families of language models: LLaMa and ChatGPT. We selected LLaMA 3.3 70B and GPT-4o as representative models of two fundamentally different paradigms in current large language model development. LLaMA 3.3 70B, a state-of-the-art open-source model developed by Meta, exemplifies the capabilities of publicly accessible, community-driven models. On the contrary, GPT-4o is part of the closed-source, commercially maintained family of models by OpenAI, offering strong performance through proprietary fine-tuning and infrastructure opti-

mizations. This comparison also enables a practical reflection on model accessibility: LLaMA models can be deployed locally and modified freely, while GPT-4o, although powerful, requires API access and incurs usage costs. Thus, our evaluation spans not only performance but also realistic deployment scenarios for researchers and educators working with low-resource languages like Italian.

In contrast to GPT-4o, that can be accessed programmatically only with the OpenAI API, LLaMA 3.3 70B was run via local inference using a GPU-enabled environment. These differing infrastructures also illustrate the practical trade-offs between accessibility, cost, and customization in real-world applications.

5. Results

In this section we present the results of the manual (RQ1 and RQ2) and automatic (RQ3) annotation tasks, pointing to sense categories and connectives which were particularly challenging for human annotators or most often misclassified by generative models.

5.1 Manual annotation task

The first research question inquired which level of agreement could be achieved by two highly trained annotators, familiar with both the Basel Model of Textuality and the sense categories of the PDTB. The task achieved an overall agreement of $\kappa = 0.785$ (see Table 1). Considering the difficulties of coding discourse phenomena in corpus-based analysis (Spooren and Degand 2010) and the language variety under consideration (student writing), we interpret this value as a substantial agreement between human annotators, in line with other projects annotated in the PDTB style and on languages other than English (Zeyrek et al. 2020).

Notwithstanding the satisfactory overall reliability, there is variation across sense categories and individual connectives. A confusion matrix grouped per sense category is available in Figure 1, while a detailed breakdown per connective form and connective sense is provided in Appendix B. Notably, disagreements gather between the Comparison:Contrast and Comparison:Concession relations, along the axis of Contingency:Cause and, more generally, in the choice between connective and non-connective uses. This is particularly evident for connectives like *ma* (English *but*), *e* (English *and*), *per* (English *for*), which are notoriously polysemic, highly frequent and operating on both the phrasal, clausal and discourse level. However, also less frequent connectives, such as *come* (English *as, like*) or *quindi* (English *therefore*) display interesting polysemy and confusion patterns. Possible annotation difficulties causing these disagreements are discussed in Section 6, contrasting them with the results obtained from the automatic annotation experiments.

The possibility to add new connectives during the annotation task was exercised in practice and resulted in the addition of the following connective forms: *a seguito di, di seguito, grazie a, in oltre, in primis, nonché, oltre che, oltre a, perchè, pero, pur*. Some of them are simply incorrect orthographic variants of connectives that were already in the LICO, such as *perchè* (*because*) and *pero* (*but*); others are forms worth entering the lexicon as they are more often than not used as connectives (e.g., *grazie a* (*thanks to*), Contingency:Cause; *nonché* (*as well as*), Expansion:Conjunction; *pur* (*even if*), Contrast:Concession).

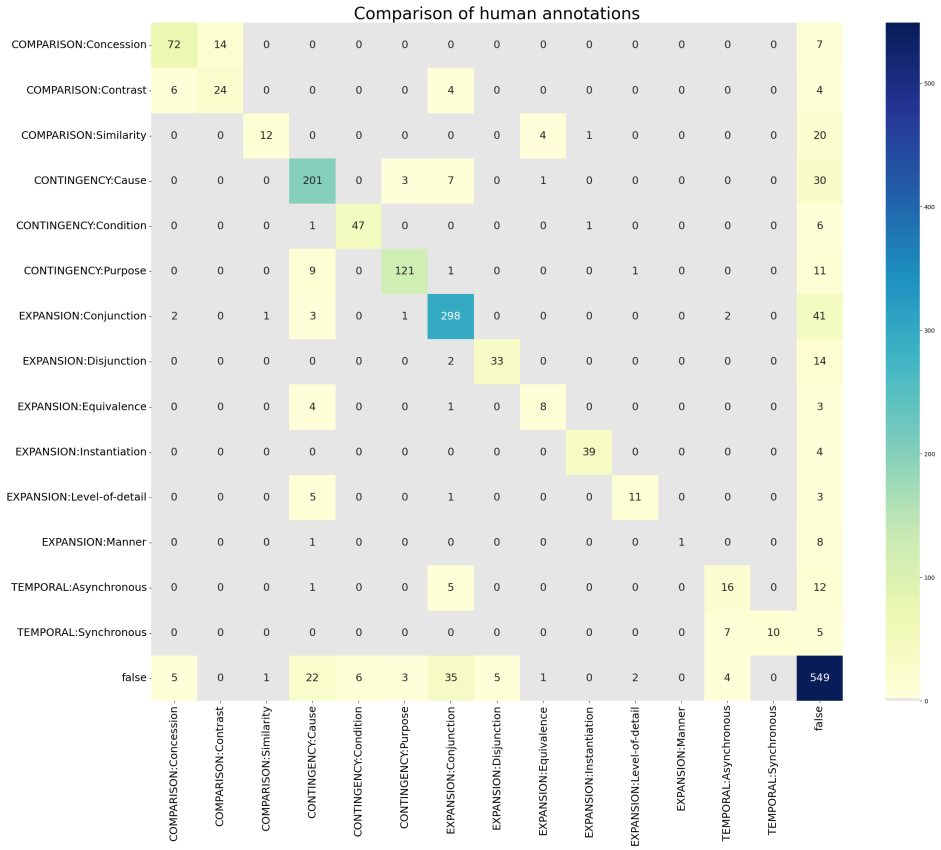


Figure 1
Confusion matrix displaying agreement patterns between two human annotators labeling connective senses. Diagonal cells (darker shading) represent perfect agreement. Off-diagonal cells reveal systematic disagreement patterns, suggesting ambiguity between senses or in distinguishing between connective and non-connective uses. The matrix demonstrates substantial overall agreement with concentrated values along the diagonal, though certain sense distinctions (e.g., within Contingency and Comparison categories) show some confusion patterns.

5.2 Experiments with LLMs

After collecting the outputs from the LLMs, we evaluate them based on precision, recall, and F1 score (micro), using as the gold standard the manual annotations of the test set for which both human annotators agreed. Table 1 reports the numerical results for the one-by-one strategy (left) and the inter-annotator agreement (Cohen's kappa) for the full-text strategy (right). The best result for the one-by-one strategy are obtained by the experiment with the LLaMa model and long prompt ($F1 = 0.521$), while for the full-text strategy the best inter-annotator agreement is between annotator 2 and GPT-4o prompted with the long prompt version ($\kappa = 0.482$). These results are slightly worse than the F1 score reached by the system proposed in Bourgonje and Demberg (2024) for the 11-way classification. Interestingly, providing the connective and sense category definitions (long prompt) increases the overall performance in terms of F1 score only

Table 1
Results on the one-by-one strategy in the different configurations (left), and inter-annotation agreement between humans and LLMs in the full-text strategy (right).

Model	Prompt	P	R	F1
llama	long	0.562	0.503	0.521
llama	short	0.507	0.584	0.515
openai	long	0.516	0.329	0.367
openai	short	0.483	0.424	0.441

openai-short	llama-long	llama-short	annotator 1	annotator 2	
0.772	0.580	0.585	0.474	0.482	openai-long
	0.538	0.613	0.435	0.434	openai-short
		0.692	0.332	0.306	llama-long
			0.298	0.288	llama-short
				0.785	annotator 1

marginally for LLaMa 3.3 70B, while it even worsens it in the case of OpenAI GPT-4o in the one-by-one scenario. Furthermore, it can be noted that prompting strategies do not alter completely the models' behavior, since highest kappa values are observed between short and long prompting strategies of the same model and not between different models prompted in the same way (see Table 1). We can conclude that exposure to labeled examples alone might be sufficient for reaching the ceiling performance. This finding can be related to the creation of annotation guidelines, which are often quite theoretical and descriptive and report just few examples. It could be the case that also for humans minimal example-centered guidelines, paired with extensive training, work equally well or even better.

Figure 2 presents the confusion matrices corresponding to the one-by-one strategy. Per-connective breakdowns are available in Appendix B. From the coarse-grained, sense-aggregated results it is already quite evident that there are systematic differences in model behaviour. Particularly, OpenAI models showed a stronger tendency for assigning non-connective usage to a variety of connectives, across almost all sense categories. Thereby, the biggest discrepancies between the the gold standard and the model annotations can be seen for the multilevel, polysemous connectives mentioned above: *e*, *per* and *o*. This result is not particularly surprising and aligns with the problems human annotator were facing for those forms.

More surprisingly, model misclassifications were also found for forms that are almost exclusively used with a connective function, both in standard language and in our student corpus (e.g., *quindi* or *infatti*, see Appendix B). Especially the GPT-4o model disproportionally misclassified such forms as non-connectives, suggesting that it might be more conservative in assigning connective status. Since the 'false' label (not a connective) was the most common one overall in the evaluation sample, this behaviour might reflect a bias for choosing the most common category, independently of the distribution of labels on a per-connective basis.

6. Discussion

Connective polysemy and the possibility for certain connective forms to operate both on a phrasal, clausal and discourse level seem to be a major source of difficulty for humans and LLMs alike. Throughout the discussion we will show how the variety

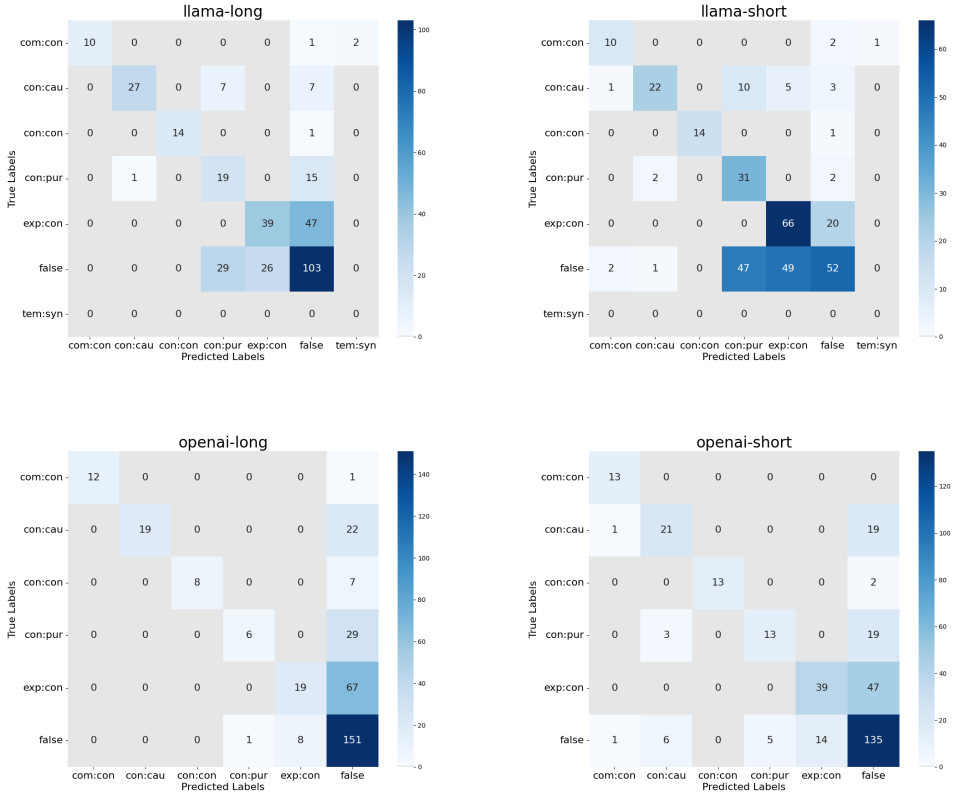


Figure 2
Confusion matrices evaluating automated discourse connective annotations against manual annotations. Results for LLaMa-long (top left), LLaMa-short (top right), GPT-4o-long (bottom left), and GPT-4o-short (bottom right). Labels represent three-letter abbreviations of PDTB sense categories (e.g., com:con = Comparison:Concession); "false" indicates non-connective classifications. Color intensity represents prediction frequency, with diagonal elements showing correct classifications.

under investigation, student writing, intersects with these sources of difficulty and might explain some of the systematic disagreements that emerged during the tasks.

6.1 Multilevel operators

The connectives *e* (*and*) and *per* (*for*) are the most common connectives in the evaluation sample. Both of them are multilevel operators and their identification according to the Basel Model of Textuality depends on subtle semantic hints in the units they link. They thus show similar confusion patterns, in particular when deciding whether they should be considered connectives or not.

Multi-level operators can operate between both discourse units and smaller units, such as phrases. Consider for example the following sentences: “I like apple and pears” vs. “I like apples and John likes pears”. The first *and* operates inside the noun phrase among units between which it is difficult to identify any discourse relation. The second

instead connects two units among which a discourse relation might hold (in this case could be Comparison:Contrast). Every discourse framework proposes different ways of segmenting the text (Hoek, Evers-Vermeul, and Sanders 2017) and different procedures for identifying connectives: from the broadest definition, which considers logical operators (*and*, *or*, *but*) always as connectives, no matter the level on which they operate and the semantics of the elements they bridge (Mauri and van der Auwera 2012), to the narrowest definition, which identifies connectives only when linking discourse units. So, depending on the definition of connective employed, only the second *and* might be counted as connective.

The Basel Model of Textuality takes an intermediate stance: the identification of connectives does not rely on the level at which the connective word operates (phrase or higher), but primarily on the semantics of the linked elements. If the linked elements are states or events, no matter their linguistic realization, the linking word is always considered a connective. For example, conjoined verbal phrases connect events, and even if they might be considered part of the same discourse unit, the coordinating conjunction is still considered a connective. The same holds for nominalizations, which might be even part of the same noun phrase. Below we display two examples from the corpus, exemplifying conjoined events when realized as verbs (1) and nouns (2).

(1) molte volte gli orari di lezioni e altro si andavano a sovrapporre con impegni di studenti, aumentando le assenze totali dei ragazzi **e** creando uno scompiglio generale per capire chi si collegherà

"Many times, class schedules and other [commitments] overlapped with students' commitments, increasing the total number of absences **and** creating general confusion as to who would be connecting"

(2) Il 1 gennaio 2020 le autorità disposero la chiusura del mercato **e** l'isolamento di coloro che presentavano segni e sintomi dell'infezione

"On 1 January 2020, the authorities ordered the closure of the market **and** the isolation of those showing signs and symptoms of infection"

To-infinitives (in Italian *per* + infinitive, Example 3) and when the prepositon *per* introduces a Cause or a Purpose (4) follow a similar line of reasoning.

(3) In data odierna ho il piacere di scriverle questa lettera **per** esprimere la mia opinione a riguardo alla sua decisione

"Today I am pleased to write this letter **to** express my opinion regarding your decision"

(4) la società che sta già soffrendo **per** i danni che sono stati arrecati dalla pandemia

"the society that is already suffering **due to** the damage caused by the pandemic"

The stance of the Basel Model of Textuality to consider states and events, in whatever way they are verbalized, as possible arguments of connectives is theoretically justified. However, it creates difficulties in annotation, particularly when encountering

grey areas such as list-like constructions, in which nouns with different semantics might be mixed (see Example 5).

(5) Le famiglie degli studenti dichiarano (per l'a.S. 2020/21) segnali di stanchezza e scarsa concentrazione (16%), problemi di socializzazione (12%) **e** ridotta capacità di seguire le lezioni (9%).

"Students' families report (for the 2020/21 school year) signs of fatigue and poor concentration (16%), socialisation problems (12%) **and** reduced ability to follow lessons (9%)."

When dealing with *per*, additional difficulties are introduced when deciding which kind of semantic relation it conveys. *Per*, when used as a connective, expresses primarily a Purpose relation, especially when used in *to*-infinitives. However, in a minority of cases it conveys a Contingency:Cause relation. Humans agree seldom on when to assign this meaning (see Example 6 for an ambiguous case) and models (and in particular LLaMa with short prompt) have a strong tendency towards almost completely disregarding this sense, assigning most instances to Purpose.

(6) Abbiamo conosciuto nuove piattaforme come Meet **per** fare le videolezione

'We have discovered new platforms like Meet **to** do [or *because of*?] videolectures'

It is no surprise that these elements are frequently disagreed upon. They are very common, polysemous and their theoretical status is still debated. More interesting instances are connectives such as *come* and *quindi*, which are less frequent, yet quite polysemous in student texts.

6.2 Core and peripheral senses

Those forms that more often than not function as connectives might be difficult because of their polysemy. However, not all senses shared by the same connective seem to trigger disagreements or misclassifications: some might be more challenging than others.

An interesting example comes from the analysis of *come* (*as, like*), a connective that can either introduce examples (Expansion:Instantiation) or metaphors and similarities (Comparison:Similarity). Most human disagreements centered on the Comparison:Similarity sense, and in particular on the construction "*come* + reporting verb" (e.g., *come detto in precedenza*, English: *as stated before*). One annotator systematically labeled these constructions as Similarity discourse relations, while the other consistently did not. This disagreement has been addressed during the curation phase, with a decision to consistently treat such constructions as expressing a Similarity sense category. When looking at models behaviour, we observe that LLaMa 3.3 chooses this category only in a minority of cases, thus aligning more closer to humans, while GPT-4o never labeled *come* with this category, rather, it didn't assign connective status to those Comparison:Similarity instances on which humans agreed, following the general trend of underlabeling also seen for other connectives. Notwithstanding the general trend, when *come* is used to express an Instantiation (i.e., it introduces an example), it seems to be not particularly confusing neither for humans nor for the two generative models.

There are then other forms, such as *perché* (*because*) or *infatti* (*indeed*), which almost always carry a certain meaning. Humans tend to show a preference for choosing the most common meaning, also when other meanings would fit best the specific occurrence under scrutiny. Additionally, in student data, variability in the use can add additional polysemy and new senses, which annotators might not expect in the controlled setting of an argumentative essay. *Quindi* is a case in point. It is a frequent connective, with relatively few core meanings. Italian dictionaries list the Contingency:Cause or maximally the Expansion:Level-of-Detail, which stands for conclusions, as possible senses. Example 7 exemplifies the causal meaning.

(7) in dad molti studenti hanno detto che durante le lezioni a distanza hanno delle difficoltà di attenzione e **quindi** faticano a memorizzare la lezione imparando solo parzialmente gli argomenti svolti.

"In distance learning, many students have said that they have difficulty concentrating during lessons and **therefore** struggle to memorise the lesson, only partially learning the topics covered"

In addition to this core meaning, both annotators found the Expansion:Equivalence sense, not listed in the dictionaries, which is probably typical of spoken language and emerges in students writing (see Example 8). Yet they rarely (2 times over 7 total instances found by the two annotators independently) agree on the cases that would be representative of this new sense.

(8) i cosiddetti "project work", ossia lavori di gruppo svolti virtualmente, **quindi** ognuno fa la sua parte e poi il vengono unite le parti formando un unico lavoro

"so-called "project work", i.e. group work carried out virtually, where everyone does their part and then the parts are combined to form a single piece of work"

This case exemplifies well how new or hybrid uses in students' writing can challenge existing sense taxonomies and annotators expectations about what is to be found in formal argumentative essays.

6.3 Contrast or Concession?

The peculiarities of student writing might also have influenced the systematic disagreements found in distinguishing Comparison:Contrast from Comparison:Concession. Within the Comparison macro-relation, distinguishing between Comparison (e.g., "John is tall, while Bob is short") and Concession relations (e.g., "John is tall, but he has never played basket") is notoriously difficult (Webber et al. 2019, 23). Even highly trained annotators couldn't assign instances to one of these categories reliably in many cases in our student texts. Consider the excerpt below (9):

(9) Questi ultimi 3 anni di scuola sono stati molto compromessi per via del Covid-19, la DaD (didattica a distanza) e la ddi (didattica digitale integrata) sono stati una novità **ma** non solo per noi studenti ma per tutto l'ambito scolastico.

> "The last three years of school have been greatly affected by COVID-19. Distance learning and integrated digital learning were new **but** not only for us students but for the entire school system."

This sentence presents some planning problems which might have added a layer of annotation difficulty for annotators. In this context, *ma* seems to elicit a Comparison:Concession discourse relation, with an implicit negated expectation, such as "you reader might think it was news just for us"[5]. Yet, the remainder of the sentence, with a second corrective *ma* softens that reading. The connective becomes superfluous (if removed completely, it does not change the discourse structure of the sentence) and it seems a leftover from a change of planning (which very often occurs in students' texts) that disrupts the syntax and is most probably a source for disagreement.

Scrutinizing the examples of disagreement, though, most cases can be traced back to the *non solo ... ma anche* construction (*not only ... but also*), with corrective sense, which is perfectly grammatical and well-formed in Italian and is particularly frequent in students essays, as well as in general language. From a purely theoretical point of view, this construction is indeed hardly classifiable as Contrast or Concession, since its main meaning is to correct the first item with the second. It shares with Contrast the absence of a negated expectation needed for Concession to be labeled as such, but it also shares with Concession the fact that the relation is between the same item and not, as in Contrast, between two items which are compared and contrasted (Izutsu 2008; Ferrato 2025). So, in absence of a specific sense category for this meaning and for this construction, probably over-used in student data, annotators could more or less freely decide for each occurrence which criteria (absence of negated expectation or absence of two items to be compared) weights the most and assign the category accordingly.

7. Conclusions

In this article, we have presented the manual and automatic annotation efforts conducted by two trained annotators and with the help of two Large Language Models, highlighting the challenges encountered in the process. Our findings not only shed light on the strengths and difficulties of both manual and automatic annotations, but also reflect broader challenges within the theoretical landscape of textual models, particularly regarding the identification and classification of discourse relations.

For future work we aim to extend the manual annotations to alternative lexicalizations and implicit relations, as one of the limitations of the current study regards the manually annotated evaluation sample, which focuses solely on explicit discourse relations, thereby restricting our findings to the simplest scenario. We also envision to refine our automatic approach, in particular in terms of prompt engineering, using the DSPy framework (Khattab et al. 2023). We acknowledge that prompt design has a significant impact on model performance. An F1 score of approximately 0.5 for the best performing experiment leaves a lot of margin for improvement. Additionally, the models we employed come with accessibility constraints: OpenAI GPT-4o require API access and associated costs, while Llama 3.3 70B demands substantial computational resources. On a more theoretical level, we would like to explore how comprehensibility

5 This interpretation might be further confirmed by the sentence following, which actually uses this discourse unit in an Comparison:Concession relation, introduced by *pure*: "Pur essendo una novità e qualcosa a cui noi studenti e professori ci dovevamo adattare secondo me la ddi dovrebbe venir estesa permanentemente per i prossimi anni scolastici"

and difficulty of a text influence inter-annotator agreement and model performances, as we have observed that peculiarities of student texts might count as a source of difficulty that intersects with other, more structural, ones. Addressing these limitations in future research could provide a more comprehensive picture of challenges in the task of shallow discourse parsing.

8. Release

To support transparency, reproducibility, and further research, all resources developed for this study are made openly available.

The software tools used for pre-processing the data and evaluating annotation results, including inter-annotator agreement metrics, are released under the GNU General Public License v3. The evaluation dataset, consisting of the manually annotated student texts with connective labels (before and after curation), is distributed under a Creative Commons Attribution (CC-BY 4.0) license. Both the code and the dataset are available on Github.[6]

References

Anuranjana, Kaveri. 2023. DiscoFlan: Instruction fine-tuning and refined text generation for discourse relation label classification. In Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes, editors, *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 22–28, Toronto, Canada, July. The Association for Computational Linguistics.

Barbagli, Alessia, Felice Dell'Orletta, Giulia Venturi, Pietro Lucisano, and Simonetta Montemagni. 2015. Il ruolo delle tecnologie del linguaggio nel monitoraggio dell'evoluzione delle abilità di scrittura: Primi risultati. *Italian Journal of Computational Linguistics*, 1(1):105–123.

Bienati, Arianna, Jennifer-Carmen Frey, Alessio Palmero Aprosio, and Noemi Facchinelli. 2023. Applicazione delle risorse disponibili per l'italiano all'annotazione automatica delle relazioni discorsive in testi scolastici: alcune implicazioni teoriche. In *LVI Congresso Internazionale della Società di Linguistica Italiana - WS3: Linguistica teorica e trattamento automatico delle lingue: verso nuove sinergie*, Torino, Italy, September.

Bourgonje, Peter and Vera Demberg. 2024. Generalizing across Languages and Domains for Discourse Relation Classification. In Tatsuya Kawahara, Vera Demberg, Stefan Ultes, Koji Inoue, Shikib Mehri, David Howcroft, and Kazunori Komatani, editors, *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 554–565, Kyoto, Japan, September. Association for Computational Linguistics.

Braud, Chloé, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes, editors, *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada, July. The Association for Computational Linguistics.

Chapados Muermans, Thomas and Leila Kosseim. 2022. A bert-based approach for multilingual discourse connective detection. In Paolo Rosso, Valerio Basile, Raquel Martínez, Elisabeth Métais, and Farid Meziane, editors, *Natural Language Processing and Information Systems*, pages 449–460, Cham. Springer International Publishing.

Cisotto, Lerida and Nazzarena Novello. 2012. La scrittura di sintesi di studenti del primo anno di scienze della formazione primaria. *Giornale Italiano della Ricerca Educativa*, 5(8):41–57.

Daumé III, Hal. 2007. Frustratingly Easy Domain Adaptation. In Annie Zaenen and Antal van den Bosch, editors, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

6 https://github.com/arianna-bienati/itaca-processing

Feltracco, Anna, Elisabetta Jezek, Bernardo Magnini, and Manfred Stede. 2016. LICO: A Lexicon of Italian Connectives. In Anna Corazza, Simonetta Montemagni, and Giovanni Semeraro, editors, *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016*, pages 141–145, Napoli, Italy, December. Accademia University Press.

Feltracco, Anna, Bernardo Magnini, and Elisabetta Jezek. 2017. Contrast-Ita Bank: A corpus for Italian Annotated with Discourse Contrast Relations. In Roberto Basili, Malvina Nissim, and Giorgio Satta, editors, *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017*, pages 159–164, Roma, Italy, December. Accademia University Press.

Ferrari, Angela. 2010. Connettivi. In *Enciclopedia dell'Italiano*. Treccani, Roma.

Ferrari, Angela. 2014. *Linguistica Del Testo. Principi, Fenomeni, Strutture*. Carocci, Roma.

Ferrari, Angela. 2021. Segnali discorsivi e connettivi. *Lingua e Stile*, 56(1):143–150.

Ferrari, Angela, Letizia Lala, and Luciano Zampese. 2021. *Le Strutture Del Testo Scritto. Teoria e Esercizi*. Carocci.

Ferrato, Elena. 2025. *La concessione in italiano: caratteristiche semantiche e sintattiche in testi scritti da giovani.* PhD thesis, University of Verona, Verona.

Hoek, Jet, Jacqueline Evers-Vermeul, and Ted J.M. Sanders. 2017. Segmenting discourse: Incorporating interpretation into segmentation? *Corpus Linguistics and Linguistic Theory*, 14(2):357–386.

Izutsu, Mitsuko Narita. 2008. Contrast, concessive, and corrective: Toward a comprehensive study of opposition relations. *Journal of Pragmatics*, 40(4):646–675.

Johannsen, Anders and Anders Søgaard. 2013. Disambiguating explicit discourse connectives without oracles. In Ruslan Mitkov and Jong C. Park, editors, *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 997–1001, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Khattab, Omar, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines, October. arXiv:2310.03714 [cs].

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico, USA, August.

Liu, Wei, Yi Fan, and Michael Strube. 2023. Hits at disrpt 2023: Discourse segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49, Toronto, Canada, July.

Magnini, Bernardo, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. I-CAB: The Italian Content Annotation Bank. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Mann, William and Sandra Thompson. 1988. Rethorical Structure Theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Mauri, Caterina and Johan van der Auwera. 2012. Connectives. In Keith Allan and Kasia Jaszczolt, editors, *Cambridge Handbook of Pragmatics*. Cambridge University Press, pages 377–402.

Metheniti, Eleni, Chloé Braud, Philippe Muller, and Laura Rivière. 2023. DisCut and DiscReT: MELODI at DISRPT 2023. In Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes, editors, *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42, Toronto, Canada, July. ACL: Association for Computational Linguistics. In conjunction with ACL 2023 and CODI 2023 workshop.

Miltsakaki, Eleni, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 2237–2240, Lisbon, Portugal, May.

Montemagni, Simonetta, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria T. Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, and Rodolfo Delmonte. 2000. The Italian Syntactic-Semantic Treebank: Architecture, Annotation,

Tools and Evaluation. In Anne Abeille, Thorsten Brants, and Hans Uszkoreit, editors, *Proceedings of the Workshop "Linguistically Interpreted Corpora", in conjunction with the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 18–27, Saarbrücken, Germany, August. International Committee on Computational Linguistics.

Palmero Aprosio, Alessio. 2021. Tint, the Swiss-Army Tool for Natural Language Processing in Italian. In Elena Cabrio, Danilo Croce, Lucia C. Passaro, and Rachele Sprugnoli, editors, *Proceedings of the Fifth Workshop on Natural Language for Artificial Intelligence (NL4AI 2021) co-located with 20th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2021)*, Online, November.

Pareti, Silvia and Irina Prodanof. 2010. Annotating Attribution Relations: Towards an Italian Discourse Treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 3566–3571, Valletta, Malta, May. European Language Resources Association (ELRA).

Pitler, Emily and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009)*, pages 13–16, Suntec, Singapore, August. Association for Computational Linguistics.

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

Prasad, Rashmi, Webber, Bonnie, Lee, Alan, and Joshi, Aravind. 2019. Penn Discourse Treebank Version 3.0.

Ruele, Michele and Elvira Zuin. 2020. Come cambia la scrittura a scuola. Technical report, IPRASE.

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

Scholman, Merel, Tianai Dong, Frances Yung, and Vera Demberg. 2021. Comparison of methods for explicit discourse connective identification across various domains. In Chloé Braud, Christian Hardmeier, Junyi Jessy Li, Annie Louis, Michael Strube, and Amir Zeldes, editors, *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 95–106, Punta Cana, Dominican Republic and Online, November. Association for Computational Linguistics.

Sechidis, Konstantinos, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the Stratification of Multi-label Data. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer.

Spooren, Wilbert and Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2):241–266.

Sprugnoli, Rachele, Sara Tonelli, Alessio Palmero Aprosio, and Giovanni Moretti. 2018. Analysing the Evolution of Students' Writing Skills and the Impact of Neo-standard Italian with the help of Computational Linguistics. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*. Accademia University Press, Torino, Italia, 10-12 December, pages 354–359.

Stede, Manfred, Tatjana Scheffler, and Amália Mendes. 2019. Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives. *Discours*, 24.

Stede, Manfred and Carla Umbach. 1998. DiMLex: A lexicon of discourse markers for text generation and understanding. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*, pages 1238–1242, Montreal, Quebec, Canada, August.

Szymański, Piotr and Tomasz Kajdanowicz. 2017. A network perspective on stratification of multi-label data. In Luís Torgo, Bartosz Krawczyk, Paula Branco, and Nuno Moniz, editors, *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA 2017) at ECML PKDD 2017*, pages 22–35, Skopje, Macedonia, September. PMLR.

Tonelli, Sara, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of Discourse Relations for Conversational Spoken Dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2084–2090, Valletta, Malta, May. European Language Resources Association (ELRA).

Webber, Bonnie. 2004. D-LTAG: Extending lexicalized TAG to discourse. *Cognitive Science*, 28(5):751–779.

Webber, Bonnie, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual.

Zanasi, Lorenzo, Arianna Bienati, Jennifer-Carmen Frey, and Chiara Vettori. 2024. Condizioni di coerenza e procedure di coesione nella scrittura scolastica: Il caso dei connettivi. In Simone Mattiola and Maja Miličević Petrović, editors, *CLUB Working Papers in Linguistics*, volume 8. CLUB – Circolo Linguistico dell'Università di Bologna, pages 131–152.
Zeyrek, Deniz, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, 54(2):587–613.

Appendices

A. Full prompts

This appendix presents the full text of the prompts used in the interaction with the Large Language Model (LLM) for the identification of connectives in Italian students' texts. Two distinct prompts were employed:

- The first prompt is more extensive and includes a concise definition of connectives, intended to guide the model's understanding of the linguistic phenomenon under investigation. It also contains some examples for each category involved in the input text.
- The second prompt is shorter and relies solely on a few-shot approach, providing the model with a series of annotated examples without an explicit definition.

Long prompt

```
In linguistica, ci sono alcuni connettivi detti COMPARISON, che hanno tre
sotto-categorie: contrast, similarity, concession.
* Contrast: at least two differences between Arg1 and Arg2 are highlighted
(es. al contrario, bensì).
* Similarity: one or more similarities between Arg1 and Arg2 are
highlighted with respect to what each argument predicates as a whole or to
some entities it mentions (es. allo stesso modo).
* Concession: a causal relation expected on the basis of one argument is
cancelled or denied by the situation described in the other (es.
prototypically tuttavia).

Altri connettivi sono di tipo TEMPORAL e possono avere due categorie:
Synchronous e Asynchronous.
* Synchronous: some degree of temporal overlap between the events
described (es. typically mentre, quando).
* Asynchronous: one event is described as preceding the other (es.
typically prima che/di, dopo).

Ci sono poi i connettivi di tipo CONTINGENCY, che possono essere: Cause,
Condition, Negative, Purpose.
* Cause: situations described in Arg1 and Arg2 are causally influenced but
are not in a conditional relation (es. typically perché, quindi).
* Condition: one argument presents a situation as unrealized (the
antecedent), which (when realized) would lead to the situation described
by the other argument (the consequent) (es. typically se, purché).
* Negative condition: one argument (the antecedent) describes a situation
presented as unrealized, which if it doesn't occur, would lead to the
```

situation described by the other argument (the consequent) (es. typically altrimenti, a meno che).
* Purpose: one argument presents an action that an AGENT undertakes with the purpose of the GOAL conveyed by the other argument being achieved (es. typically affinché).

Esistono anche i connettivi di tipo EXPANSION, che possono essere: Conjunction, Disjunction, Equivalence, Instantiation.
* Conjunction: both arguments bear the same relation to some other situation evoked in the discourse. It indicates that the two arguments make the same contribution with respect to that situation or contribute to it together. It differs from most other relations in that the arguments don't directly relate to each other, but to this other situation (es. prototypically e, in più at the start of a sentence).
* Disjunction: two arguments are presented as alternatives, with either one or both holding. As with Conjunction, Disjunction is used when both its arguments bear the same relation to some other situation evoked in the discourse, making a similar contribution with respect to that situation. While the arguments also relate to each other as alternatives (with one or both holding), they also both relate in the same way to this other situation (es. typically o, oppure).
* Equivalence: both arguments are taken to describe the same situation, but from different perspectives (es. typically cioè).
* Instantiation: one argument describes a situation as holding in a set of circumstances, while the other argument describes one or more of those circumstances (es. typically ad/per esempio).

Infine, alcune congiunzioni o avverbi non sono considerati connettivi.
* Non vanno considerati connettivi elementi che, pur essendo parole grammaticali, invariabili, non indicano relazioni logico-argomentative: si pensi paradigmaticamente agli introduttori delle subordinate completive (che, di ecc.) e a quelli delle subordinate relative.
* Non vanno considerate connettivi quelle espressioni che, pur essendo associate a una relazione logico-argomentativa, sono morfologicamente variabili: da ciò discende che, per questo fatto, la conseguenza è che, la causa? ecc. In quest'ultimo caso, si può parlare di para-connettivi.

Short (common) prompt

Nel compito che andrai a svolgere, occorre individuare i connettivi presenti in una frase, specificandone la tipologia.

Di seguito, alcuni consigli su come svolgere il compito.
Innanzi tutto, nella frase indicata sono state pre-annotate alcune espressioni che potrebbero essere connettivi.
Tali espressioni sono: [EXPRESSIONS].

[EXAMPLES]

Questa è la frase su cui dovrai lavorare:
[SENTENCE]

Per ciascuna espressione che ritieni essere un connettivo, indica la tipologia e la sottocategoria.
Non fornire altre informazioni né motivazioni, solo la risposta.

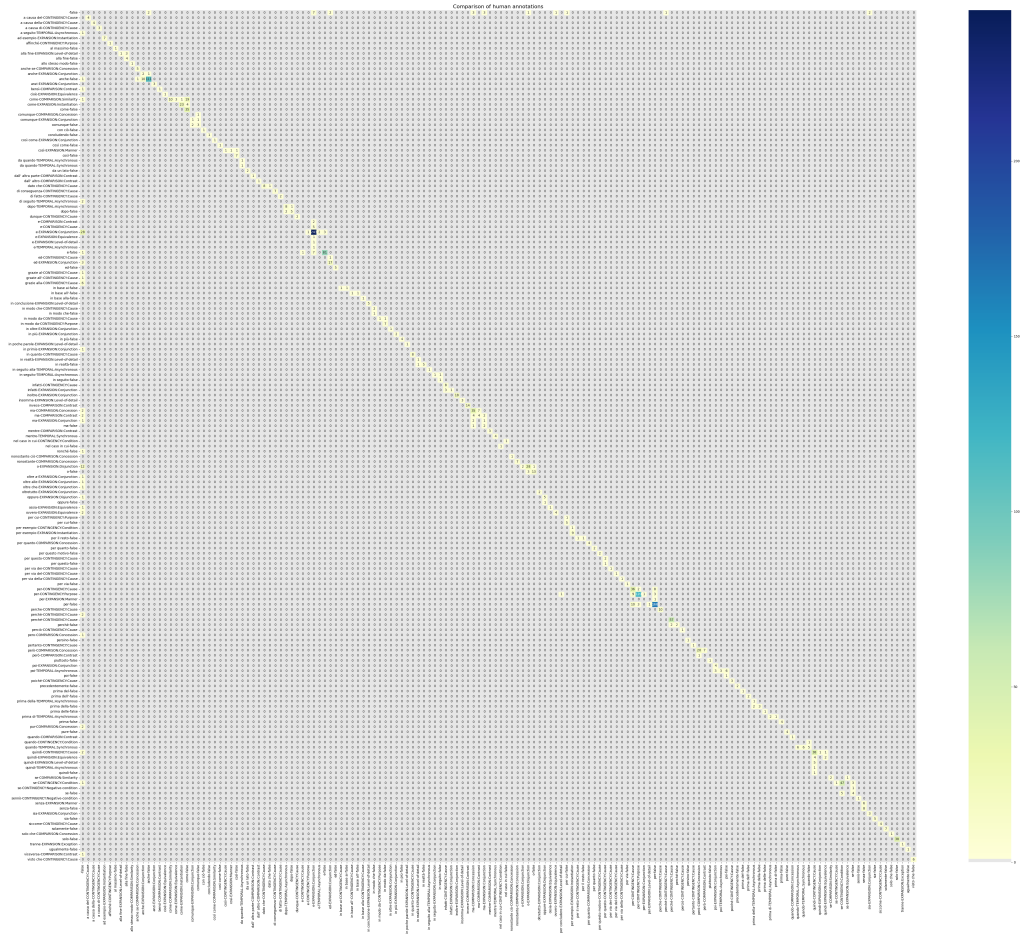
Per rispondere, scrivi un elenco puntato in cui ogni punto corrisponde a un connettivo.
Puoi ignorare le espressioni che non sono connettivi.

B. Per-connective breakdowns

This appendix contains additional illustrations showing individual connective patterns.

Manual annotation task

For the manual annotation task, we plot a confusion matrix, with connectives and respective senses as labels. The visualization helps to detect difficult connectives that brought about confusion patterns between the two human annotators. The confusion matrix is also available online at `https://github.com/arianna-bienati/itaca-processing/blob/main/img/assurdo-plot.png`, to give the possibility to the interested reader to zoom in the single connectives/sense categories.



Automatic annotation task

For the automatic annotation task, we opted for a barplot visualization, faceted by connective. On the x axis are given the senses each connective was assigned to. On

the y axis counts are grouped by the source of annotations (human vs. LLMs in the different experimental conditions). The grouped barplot helps to see the different label distributions per annotation source and per connective. It helps detect some common trends between model families and prompt strategies with respect to true (human) distributions.