

# **Applicazione delle risorse disponibili per l'italiano all'annotazione automatica delle relazioni discorsive in testi scolastici:**

Alcune implicazioni teoriche

Arianna Bienati\*, Jennifer-Carmen Frey\*, Alessio Palmero Aprosio†, Noemi Facchinelli+

\*Eurac Research, †Fondazione Bruno Kessler, +Università degli Studi di Siena

# Outline

What are we going to talk about?

## **annotazione automatica delle relazioni discorsive**

=

Discourse parsing

Aim: parse the discourse structure of a text  
automatically according to a “discourse grammar” (e.g.,  
DLTAG > PDTB, Webber 2004; Prasad et al. 2008; Webber et al. 2019)

## **risorse disponibili per l'italiano**



- Manually annotated resources  
(Feltracco et al. 2017; Pareti & Prodanof 2010; Tonelli et al. 2010)
- Lexica (Feltracco et al. 2016; Stede et al. 2019)

Alcune implicazioni teoriche

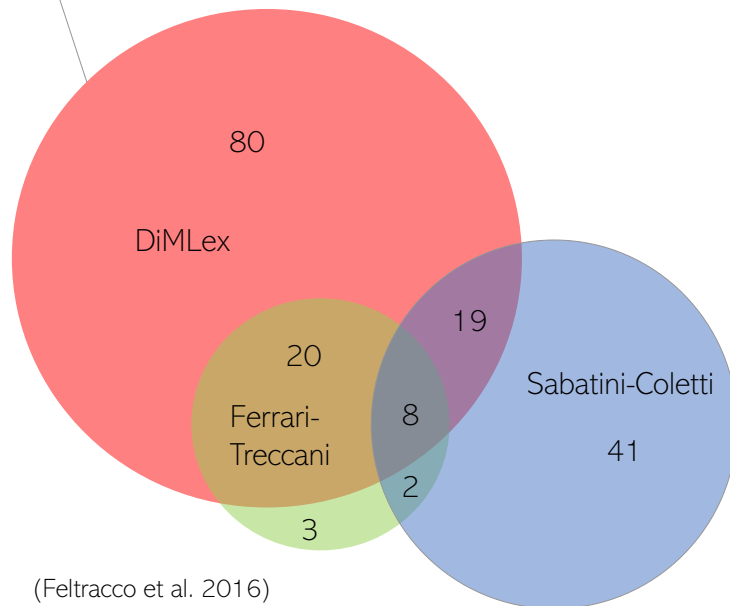
Segmentation of the text

Definition of connective

# LICO: A Lexicon of Italian Connectives

## LICO in short (Feltracco et al. 2016)

- 173 Italian connectives
- Info about: variants, syntactic categories, discourse semantics, usage examples, synonyms in the German DiMLex
- Definition of 'connective' by Ferrari (2010)
- Composed of:
  - Ferrari's (2010) list of connectives (33) in Enciclopedia Treccani
  - Sabatini-Coletti «congiunzioni testuali» (70)
  - German DiMLex equivalents (127). Some of the connectives found in each source overlap with other sources.



## a causa di

Italian  
LICO c1

**Synonyms:** aufgrund dimlex k19, aufgrund dessen, dass dimlex k21, wegen dimlex k160

**prep**

**CONTINGENCY:**Cause:reason

**Example:** *A causa del maltempo il St. Gotthard è rimasto chiuso.*

**Example:** *Una chiusura a causa del maltempo verrà presa in considerazione.*

## Positive definition

Formal  
criterion

+

Functional  
criterion

«Il termine connettivo indica 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»

(Ferrari 2021: 145)

## Negative definition

«Non vanno considerati connettivi:

- elementi che, pur essendo parole grammaticali invariabili, non indicano relazioni logico-argomentative → **no relative pronouns & complementizers**
- le preposizioni che indicano entità semantiche di primo grado, come persone, animali, cose, astrazioni → **no prep. + first order entities**
- quelle espressioni che, pur essendo associate a una relazione logico-argomentativa, sono morfologicamente variabili» → **no morphologically variable expressions:** e.g., *da ciò discende che, per questo fatto*

(Ferrari 2021: 146)

Definition of 'connective' (Ferrari, 2010; 2014; 2021)

# Bridging theory and practice

Is LICO consistent with Ferrari's (2010) definition of 'connective'?

"In our work, we partly drop the invariability criteria; we do not include forms which exhibit morphological inflection or conjugation"

(Feltracco et al. 2016)

"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".

(Feltracco et al. 2016)

<b>per questa ragione – per queste ragioni</b> <b>per [tutte] queste ragioni</b> <b>Synonyms:</b> dadurch <small>dimlex k32</small> , daher <small>dimlex k39</small> , inolgedessen <small>dimlex k92</small>	Italian LICO c130
cco CONTINGENCY:Cause:result	
<b>per questo motivo – per questi motivi?</b> <b>per [tutti] questi motivi</b> <b>Synonyms:</b> daher <small>dimlex k39</small>	Italian LICO c132
cco CONTINGENCY:Cause:result <b>Example:</b> Non aveva ancora mangiato. Per questo motivo si sentiva indebolita.	

<b>ad esempio</b>	Italian LICO c9
cco EXPANSION:Instantiation	Variants ad esempio <b>c</b> per esempio
<b>a dire la verità</b> <b>a dire [tutta] la verità</b>	Italian LICO c4
adv COMPARISON:Concession:Arg2-as-denier	Variants a dire la verità <b>c</b> in verità per la verità

LICO admits a 'certain' *paradigmatic variability*.

- Which degree of paradigmatic variability is admitted?
  - And at which degree of variability the form starts to be counted as a different entry and not a variant anymore?

Is *syntagmatic variability* admitted?

# Research questions

What do we want to achieve?

## Research questions

- RQ1: How do methodological choices regarding the segmentation of discourse units and the definition of connective influence the reconstruction and analysis of the discourse structure of students' texts?
- RQ2: To what extent do these choices influence the performance of an automatic string-matching approach for connective detection based on the LICO?

# Data & methods

How do we want to achieve it?

## Research questions

RQ1: How do methodological choices regarding the segmentation of discourse units and the definition of connective influence the reconstruction and analysis of the discourse structure of students' texts?

RQ2: To what extent do these choices influence the performance of an automatic string-matching approach for connective detection based on the LICO?

Data: two datasets of Italian student writing of different ages and proficiencies.

### LEONIDE (Glaznieks et al. 2022)

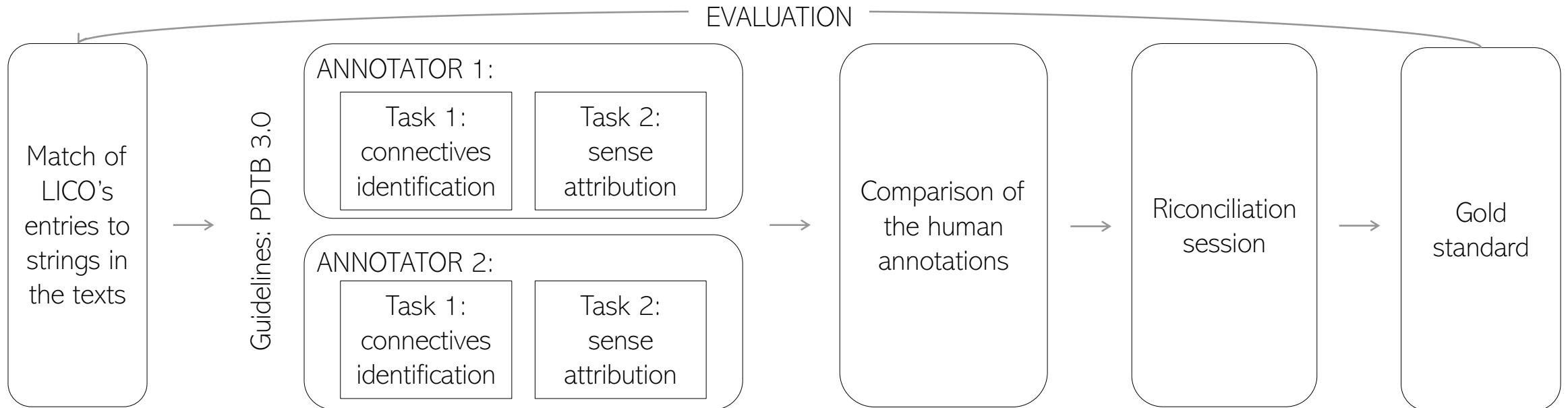
- L1 & L2 users
- Grade 6-8
- 18 picture stories and opinion texts
- 2015 tokens

### ITACA

- L1 users
- Grade 12
- 12 argumentative essays
- 7661 tokens

## Methods: human vs. automatic annotation

### EVALUATION



# Results

Overall performance of the automatic annotation in the identification of connectives

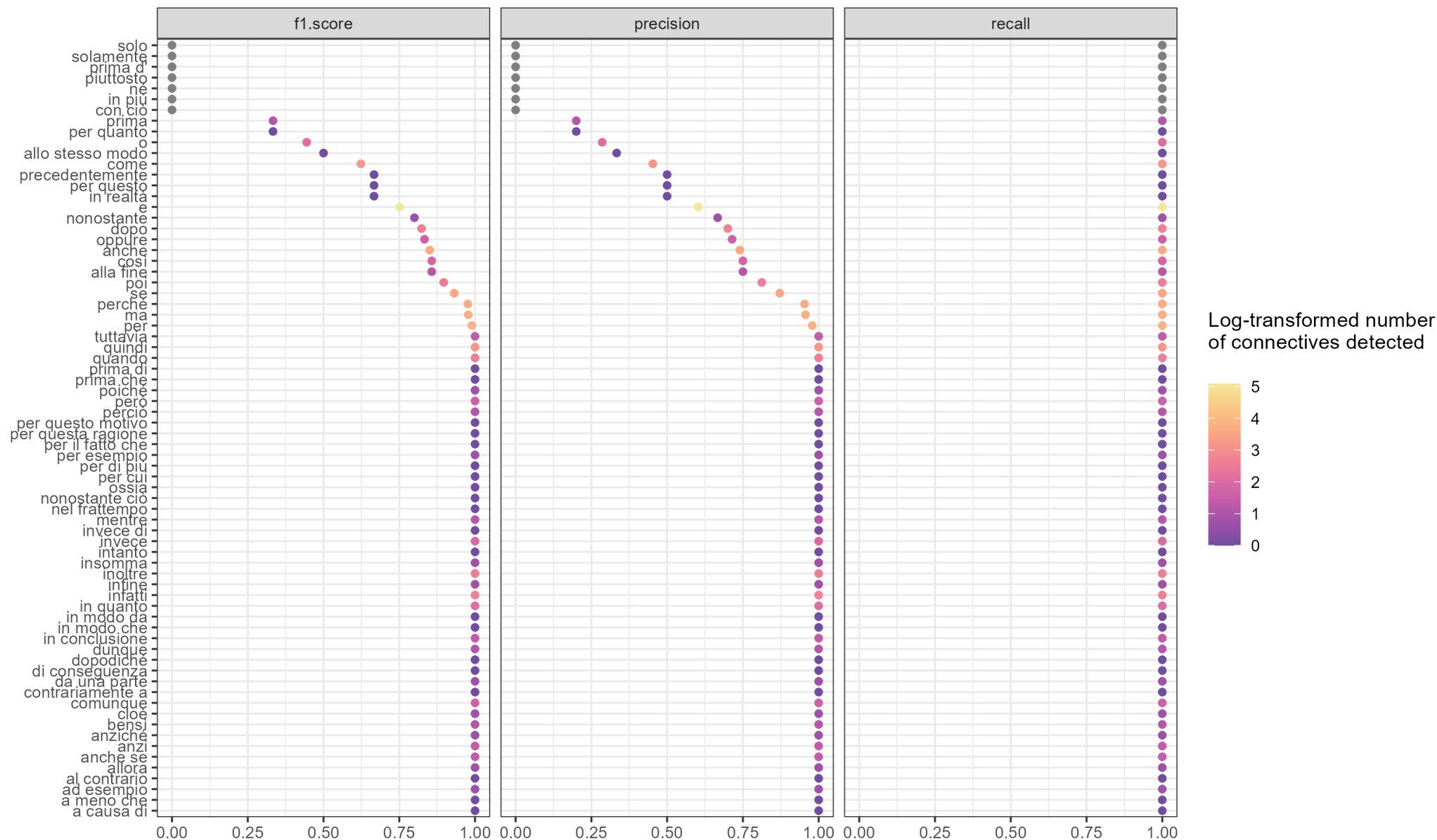
## False negatives:

- Orthographical variants: **ed, od**
- Spelling: **ho (o), ou (o), è (e), pero (però), per quand**
- Variants of *in conclusione*: **per concludere, come conclusione, concludendo**
- Morphosyntactic variants: **per quanti (per quanto)?** (es. "per quanti aspetti positivi si possano trovare della DaD, non penso comunque che questa possa essere migliore della scuola in presenza"), **al contempo (nel contempo)**
- Forms present in connective lexicon of other languages: **senza che, senza (VP inf)**
- Analogous forms: **dopo che (prima che exists)**
- Synonyms of existing forms: **ovvero (ossia), al fine di (a tal fine), per via di/dei/del/delle (a causa di), in aggiunta (in più)**
- Varia: **seppur/seppure, ora?** (es. se prima la scuola era un godere e un soffrire con gli altri, ora è diventato un isolamento sociale), **dato che, per (un motivo)?** (es. lo sostengo che in Alto-Adige sia molto importante lo studio del tedesco per un motivo, cioè quello che sapendolo e dopo aver studiato il tedesco si ha molte opportunità di lavoro sia in Alto-Adige sia all'estero).



# Results

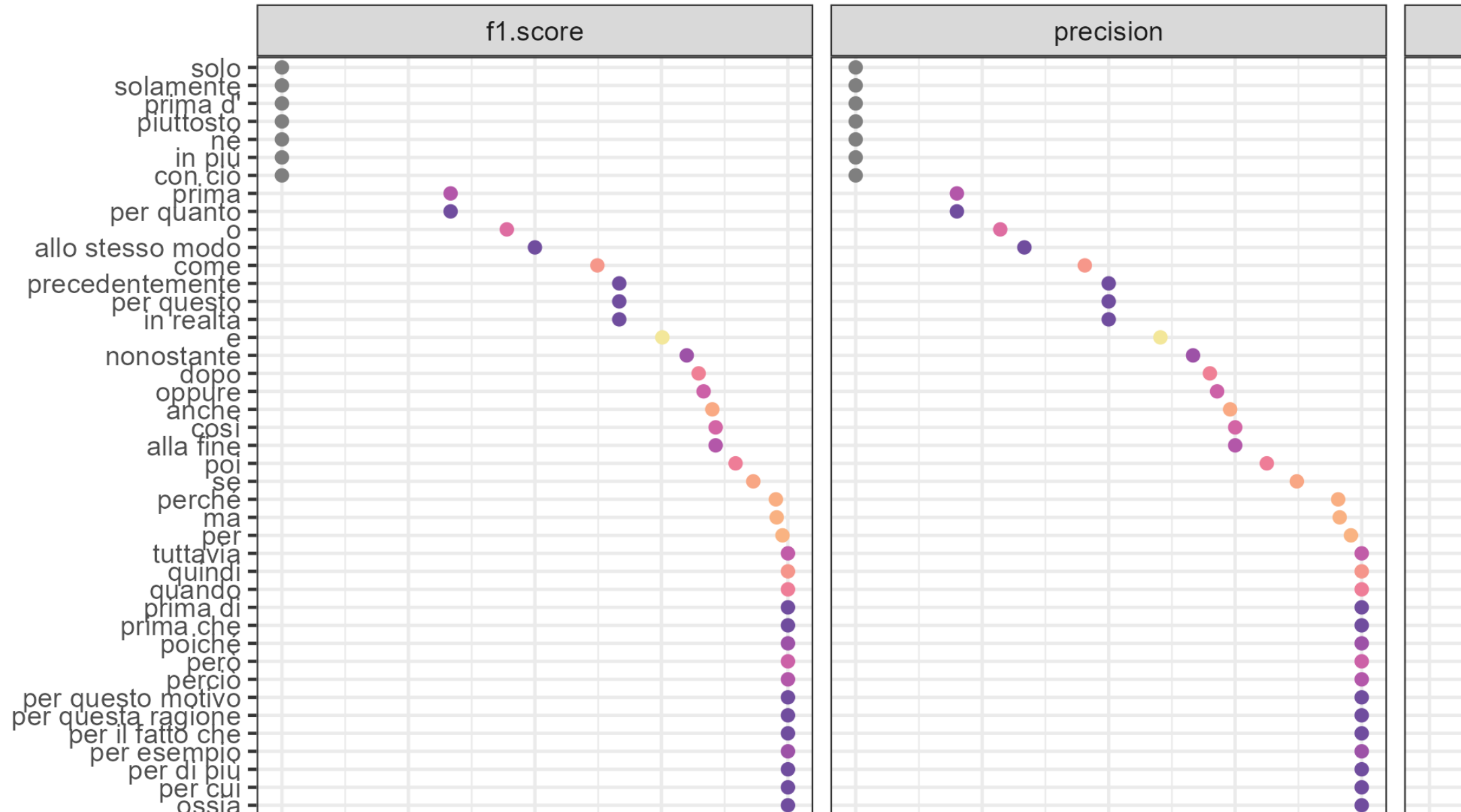
F1 , precision and recall of the automatic annotation per connective





# Results

Automatic annotation vs. human annotations – True and false positives



# Discussion

Some connective forms show low precision because of:

## Multicategoriality

The same form can perform different syntactic functions depending on the context in which it appears.

Es. *solo* (De Mauro):

- avv. *sono solo due anni che lo conosco*;
- cong. *È molto bello, solo non è il mio genere*.

Es. *prima* (De Mauro):

- avv. *Ci siamo visti prima*;
- agg. *l'ora prima*;
- ma anche cong. (LICO): *Domani vado in Spagna. Prima c'è ancora molto da fare*.

Es. *perché* (De Mauro):

- avv. in interrogative dirette o indirette
- cong. *prendi l'ombrello perché potrebbe piovere*
- s. m. inv. *spiegami il perché del tuo comportamento*

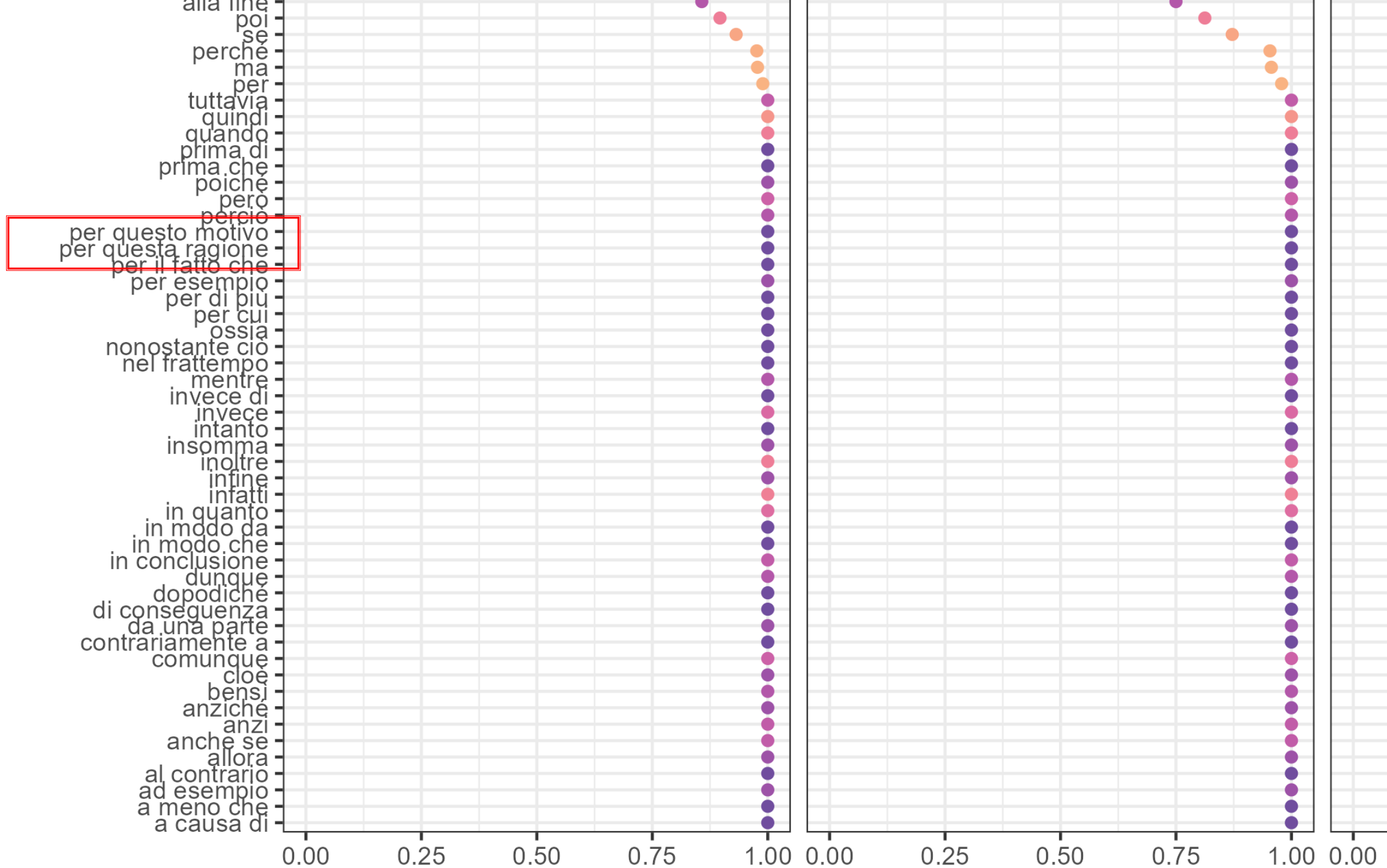
## Multilevel operators

The same form performs the same function at different levels of the discourse/syntactic structure.

Es. *e, né*:

- Discourse level (inter- and intra-sentential): *Ormai l'essere umano ha vissuto sulla terra più di duecentocinquantamila anni e col passare del tempo si è evoluto e con lui le varie tecnologie che l'uomo ha utilizzato e utilizza tutt'ora per sopravvivere*.
- Sentence level (between PPs): *in un futuro prossimo essa potrebbe non essere più in grado di provvedere né ai nostri bisogni né a quelli della flora e della fauna*.

→ The low precision of coordinate conjunctions (*e, o, ma*) depends heavily on the segmentation of the text: what do we consider as a possible argument of a connective?



# Discussion

Some connective forms show low precision because of:

## Multicategoriality

The same form can perform different syntactic functions depending on the context in which it appears.

## Multilevel operators

The same form performs the same function at different levels of the discourse/syntactic structure.

Reliability of variable composite forms vs. plain conjunctions:

“Anzitutto, all’interno dell’ovvio rapporto non biunivoco tra forme e funzioni, vi è un massimo di ambiguità nelle forme meno esplicite e morfologicamente più semplici, come e, e al capo opposto invece una totale trasparenza nelle forme più complesse ed esplicite”. (Berretta 1984: 245).

## Possible changes to the LICO:

Delete forms that more often than not perform other functions instead of the connective one?

## Possible improvements to the method:

Add rules based on the PoS-tagging to control (partially) for multicategoriality and rules based on dependency parsing to reduce the error on coordinate conjunctions.

Adding/keeping forms that might be composite and variable but clearly convey a specific discourse relation (and are reliably detectable)?

- Keeping “primary” and “secondary” connectives apart (Danlos, 2018)?
- How to check formal and functional criteria in a systematic/empirical way (excluding linguistic introspection) (e.g., querying corpora to find out how forms occur)?

## Still some questions remain...

- What is the merit of an only functional vs. a functional/formal criterion of defining connectives?
- Quantitative a-theoretic approach for defining a connective (Tamburini et al. 2002)?

# Grazie per l'attenzione

[arianna.bienati@eurac.edu](mailto:arianna.bienati@eurac.edu)

[jennifer-carmen.frey@eurac.edu](mailto:jennifer-carmen.frey@eurac.edu)

[aproso@fbk.eu](mailto:aproso@fbk.eu)

[n.facchinelli@student.unisi.it](mailto:n.facchinelli@student.unisi.it)

# References

- Danlos, L. (2018). Discourse and Lexicons: Lexemes, MWEs, Grammatical Constructions and Compositional Word Combinations to Signal Discourse Relations. *Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), Workshop at Coling 2018*. <https://hal.science/hal-02069442>
- Feltracco, A., Jezek, E., Magnini, B., & Stede, M. (2016). LICO: A Lexicon of Italian Connectives. In A. Corazza, S. Montemagni, & G. Semeraro (Eds.), *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016* (pp. 141–145). Accademia University Press. <https://doi.org/10.4000/books.aaccademia.1770>
- Feltracco, A., Magnini, B., & Jezek, E. (2017). Contrast-Ita Bank: A corpus for Italian Annotated with Discourse Contrast Relations. In R. Basili, M. Nissim, & G. Satta (Eds.), *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017* (pp. 159–164). Accademia University Press. <https://doi.org/10.4000/books.aaccademia.2407>
- Ferrari, A. (2010). Connettivi. In *Enciclopedia dell'Italiano*. Treccani. [https://www.treccani.it/enciclopedia/connettivi\\_\(Enciclopedia-dell'Italiano\)](https://www.treccani.it/enciclopedia/connettivi_(Enciclopedia-dell'Italiano))
- Ferrari, A. (2014). *Linguistica del testo. Principi, fenomeni, strutture*. Carocci.
- Ferrari, A. (2021). Segnali discorsivi e connettivi. *Lingua e Stile*, *LVI*(1), 143–150. <https://doi.org/10.1417/100904>
- Glaznieks, A., Frey, J.-C., Stopfner, M., Zanasi, L., & Nicolas, L. (2022). Leonide: A longitudinal trilingual corpus of young learners of Italian, German and English. *International Journal of Learner Corpus Research*, *8*(1), 97–120. <https://doi.org/10.1075/ijlcr.21004.gla>
- Pareti, S., & Prodanof, I. (2010). Annotating Attribution Relations: Towards an Italian Discourse Treebank. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 3566–3571.
- Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse Treebank 2.0. *Proceedings of the 6th International Conference of Language Resources and Evaluation (LREC 2008)*, 2961–2968.
- Stede, M., Scheffler, T., & Mendes, A. (2019). Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives. *Discours*, *24*. <https://doi.org/10.4000/discours.10098>
- Tamburini, F., De Santis, C., & Zamuner, E. (2002). Identifying Phrasal Connectives in Italian Using Quantitative Methods. In S. Nuccorini (Ed.), *Phrases and Phraseology—Data and Descriptions*. Peter Lang. <https://philarchive.org/rec/ZAMIPC>
- Tonelli, S., Riccardi, G., Prasad, R., & Joshi, A. (2010). Annotation of Discourse Relations for Conversational Spoken Dialogs. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2084–2090.
- Webber, B. (2004). D-LTAG: Extending lexicalized TAG to discourse. *Cognitive Science*, *28*(5), 751–779. <https://doi.org/10.1016/j.cogsci.2004.04.002>
- Webber, B., Prasad, R., Lee, A., & Joshi, A. (2019). *The Penn Discourse Treebank 3.0 Annotation Manual*.

# Results

Automatic annotation vs. human annotations – True and false positives

Ho paura che con un altro pianeta l'uomo possa **solo** distruggerlo

f1.score

precision

La natura non può fornire risorse **per** sempre **e** **prima** o **poi** giungerà ad una fine

Le nostre conoscenze didattiche si sono fermate a **prima** della quarantena

Si è molto discusso dell'utilità dell'utilizzo della DAD, quindi della Didattica a Distanza, **come** mezzo d'insegnamento alternativo

Il tempo passato davanti al computer per motivi scolastici ci spinge a ricercare nel tempo libero forme di intrattenimento alternativo **come** sport o passioni che non implicino l'utilizzo di dispositivi elettronici.

Utilizzare internet non è **così** bello **come** sembra

e questo causerebbe un grande problema **per** il pianeta ospite

**Dopo** una lunga camminata nei boschi, si fermarono **per** fare una breve pausa...

Seppur sia vero che, tante ore allo schermo **per** le lezioni ti fanno passare la voglia di sfruttare il tuo tempo libero allo stesso modo ...

Molte invenzioni tecnologiche che oggi sono di uso quotidiano, **anche** un tempo erano considerate tali

Certi scienziati piuttosto famosi cercano di convincere la gente che sia meglio lasciare la terra ... e sostengono **anche** che non ci sia alcun male esplorare nuovi pianeti

solo  
solamente  
prima d'  
con, o  
per quanto  
o  
allo stesso modo  
come  
precedentemente  
per questo  
in realtà  
e  
nonostante  
dopo  
oppure  
anche  
così  
alla fine  
poi  
se  
perché  
ma  
per  
tuttavia  
quindi  
quando  
prima di  
prima che  
poiché  
però  
perciò  
per questo motivo  
per questa ragione  
per il fatto che  
per esempio  
per di più  
per cui  
ossia