Workshop Discourse studies and linguistic data science: Addressing challenges in interoperability, multilinguality and linguistic data processing DiSLiDaS 2022

ISO-DR-core plugs into ISO-dialogue acts for a crosslinguistic taxonomy of discourse markers

Purificação Silvano & Mariana Damova

University of Porto & Mozajka Ltd
msilvano@letras.up.pt & mariana.damova@mozajka.co

Nexus Linguarum

U.PORTO
FLUP FACULDADE DE LETRAS
UNIVERSIDADE DO PORTO

MOZAIKA

Background and Motivation

- **Discourse markers:**

  - largely studied in different languages (e.g. Schiffrin (1987), Knot & Dale (1993), Fraser (1996),Taboada (2006), Silvano (2010), Das (2014), Mendes et al. (2018), Stede et al. (2019) either independently or in relation with other issues;

  ⇩

  - their relevance in discourse interpretation;
  - their complexity regarding their multifunctional nature.

Background and Motivation

- **Discourse markers**

↓

Several taxonomies within different theoretical frameworks, some language independent, others - language specific, many associated to discourse relations taxonomies (eg. Mann & Thompson (1988), Sanders et al. (1992), Asher & Lascarides (2003), Prasad et al. (2008), Zeyrek et al. (2018)), and most directed to written discourse (cf. eg. for spoken discourse Gónzalez (2005), Crible (2014) Mascher & Schiffrin (2015)).

Background and Motivation

- **Discourse markers**

  Some efforts to reconcile different taxonomies and/or to propose an overarching model for DM annotation, eg.

  - Petukhova & Bunt (2009), ISO 24617-2;
  - Prasad & Bunt (2015), Bunt & Prasad (2016), ISO 24617-8;
  - Crible (2014); Crible & Degand (2019).

Background and Motivation

Discourse markers

- Overall,
  - some taxonomies can be used to annotate the meaning of discourse markers, but only a few specifically designed for that purpose.
  - none attempts at using ISO standards that can capture both their semantic and pragmatic meaning.
  - most DM oriented taxonomies lack a wide-range application to corpora across languages, genres and types of discourse to test their reliability and comprehensiveness

Our Purpose

- Propose a comprehensive interoperable Discourse Markers taxonomy able to represent not only the semantic meaning of discourse markers but also their pragmatic meaning.

- Determine its reliability by applying it to a sample of a multilingual dataset.

Our proposal

➢ Assumptions

✓ Discourse markers

- subsume words or expressions that link utterances and play different pragmatic functions (eg. Schriffin (1987), Fraser (2009), Crible (2014), Crible & Dagand (2019)) - connectives (*as a consequence*, *on one hand*) and pragmatic particles (*you know*, *I mean*).

- are multifunctional: eg. Hovy (1995): DM convey semantic information, interpersonal purpose, and rhetorical relation.
  can have multiple meanings simultaneously (Petukhova & Bunt (2009))

Our proposal

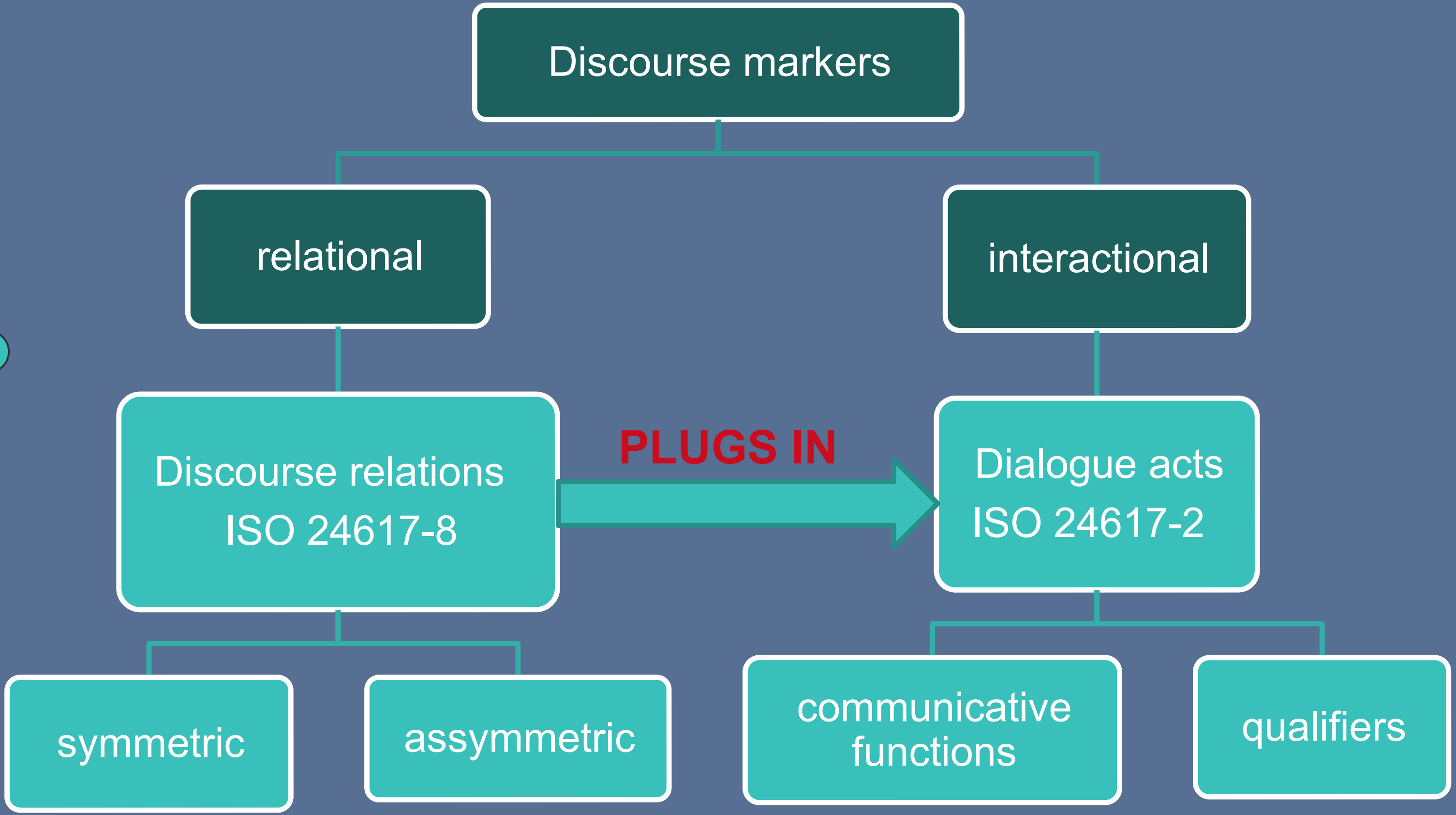
1) It turns out that rarely do we practice under the types of conditions we’re actually going to perform under, and **as a result**, when all eyes are on us, we sometimes flub our performance. **Semantic value**

2) (Applause) Lakshmi Pratury: Just stay for a second. Just stay here for a second. (Applause) **You know**, when I heard Simon’s – please sit down; I just want to talk to him for a second. **Pragmatic value**

Our proposal

3) Instead, so far, the measurements coming from the LHC show no signs of new particles or unexpected phenomena. **Of course**, the verdict is not definitive. **Semantic and pragmatic values**

Our proposal



Discourse markers

relational

interactional

Discourse relations
ISO 24617-8

PLUGS IN

Dialogue acts
ISO 24617-2

symmetric

assymmetric

communicative functions

qualifiers

Our proposal

Discourse Relations			
Asymmetric	Semantic Role		Symmetric
	Arg 1	Arg 2	
Cause	result	reason	Conjunction
Expansion	narrative	expander	Contrast
Asynchrony	before	after	Synchrony
Concession	expectation raiser	expectation-denier	Similarity
Elaboration	broad	specific	Disjunction
Exemplification	set	instance	Restatement
Manner	achievement	means	
Condition	consequent	antecedent	
Negative Condition	consequent	negated-antecedent	
Purpose	enablement	goal	
Exception	regular	exclusion	
Substitution	disfavoured-alternative	favoured-alternative	

Communicative functions		Qualifiers
General	Dimension-specific	
checkQuestion	autoPositive	conditional/
inform	autoNegative	unconditional
agreement	alloPositive	certain/uncertain
disagreement	alloNegative	positive/ negative
correction	feedbackElicitation	
answer	stalling	
confirm	pausing	
disconfirm	interactionStructuring	
offer	opening	
promise	topicShift	
addressRequest	selfError	
acceptRequest	retraction	
declineRequest	selfCorrection	
addressSuggest	initGreeting	
acceptSuggest	initSelfIntroduction	
declineSuggest	apology	
request	thanking	
instruct	initGoodbye	
suggest	compliment	
addressOffer	congratulation	
acceptOffer	sympathyExpression	
declineOffer	contactCheck	

Our proposal

4) It turns out that rarely do we practice under the types of conditions we’re actually going to perform under, and **as a result**, when all eyes are on us, we sometimes flub our performance. **Cause**

5) Ah, earth’s oceans. They are beautiful, inspiring, life-sustaining. They are also, as you’re probably quite aware, more or less screwed. In the Seychelles, **for example**, human activities and climate change have left corals bleached. Overfishing has caused fish stocks to plummet. **Exemplification**

Our proposal

6) (Applause) Lakshmi Pratury: Just stay for a second. Just stay here for a second. (Applause) **You know**, when I heard Simon's – please sit down; I just want to talk to him for a second . **Opening**

7) And that is, there is a sudden emergence and rapid spread of a number of skills that are unique to human beings like tool use, the use of fire, the use of shelters, and, **of course**, language, and the ability to read somebody else's mind and interpret that person's behavior. **Confirm/ Certain**

Our proposal

8) Instead, so far, the measurements coming from the LHC show no signs of new particles or unexpected phenomena. **Of course**, the verdict is not definitive. **Expansion/ Confirm Certain**

The proof of concept: experiment

- Aim: determine the reliability and coverage of the proposed taxonomy

- Description of the experiment

  - Dataset: 165 multiword discourse makers occurrences in 3 languages, English, European Portuguese and Bulgarian.
  - Source: TED Talk transcripts
  - Annotation procedure:
    - English as baseline
    - Annotation manual
    - Annotators of EP and BL – native speakers

Discourse markers meaning	English DM	Portuguese DM	Bulgarian DM
Exemplification	for example, for instance	por exemplo	например
Elaboration	in particular, to sum up	em suma (*in sum*)	Особено (*especially*), в частност (*in particular*)
Synchrony	so far	até agora (*until now*)	до сега (*until now*)
Contrast	on the one hand	por um lado	от една стра- на
Concession	on the other hand	por outro lado	от друга страна
Conjunction	on the other hand	por outro lado	от друга страна
Restatement	in other words, I mean	por outras palavras, noutras palavras, isto é (*this is*)	с други думи (*in other words*)
Cause	as a result	como resultado, como consequência (*as a consequence*)	в резултат
Expansion	in fact, this is, that	de facto, ou seja,	всъщност (*in fact*)

The proof of concept: results

Discourse markers meaning	English DM	Portuguese DM	Bulgarian DM
CheckQuestion	you know		знаеш ли (*you know*), знаете ли (*do you know*)
Confirm	of course, in fact	claro, de facto, na verdade (*in true*)	разбира се (*of course*)
Opening	you know	sabem	знаеш ли (*you know*), знаете ли (*do you know*)
AlloPositive	you see		виждаш ли (*can you see*)

Conclusion

- **Our proposal:**

- ✓ specifically designed to codify the meaning of DM;
- ✓ the two dimensions, semantic and pragmatic, are featured by values that are specific to those dimensions (and not generic);
- ✓ the dimensions-oriented values properly account for the role or roles each DM can play in discourse;
- ✓ being the values extracted from parts of ISO 24617, tried out in different genres and text modalities and languages, grants our proposal reliability and allows for interoperability.

Future steps

1st phase – Stabilize the taxonomy:

- Add more discourse relations to account for pertinent distinctions of meaning;
- Apply the taxonomy to a larger dataset both composed of monologues and dialogues;
- Define a smaller set of relevant communicative functions taking in consideration their occurrence on the corpora.

2nd phase – Large–scale annotation:

- Annotation of the complete corpus using inter-annotator agreement.

3rd phase – Develop an empirical-based multilingual lexicon of discourse markers to be used as LLOD.

References

Language resource management-semantic annotation framework (semaf) - part 8 - semantic relations in discourse,core annotation schema (dr-core). Standard, Geneva, CH (2016)

Language resource management-semantic annotation framework (semaf) - part 2 - dialogue acts. Standard, Geneva, CH (2020)

Asher, N., Asher, N.M., Lascarides, A.: Logics of Conversation. Cambridge University Press (2003)

Benamara, F., Taboada, M.: Mapping different rhetorical relation annotations: A proposal. In: Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics. pp. 147–152. Association for Computational Lin- guistics, Denver, Colorado (Jun 2015). https://doi.org/10.18653/v1/S15-1016, https://aclanthology.org/S15-1016

Bunt, H. (2000). Dynamic Interpretation and Dialogue Theory. M.M. Taylor, D.G. Bouwhuis and F. Neel (eds.) The Structure of Multimodal Dialogue, Vol 2., Amsterdam: John Benjamins, pp. 139–166.

Bunt, H.: Plug-ins for content annotation of dialogue acts (2019)

Bunt, H., Petukhova, V., Gilmartin, E., Pelachaud, C., Fang, A., Keizer, S., Prevot, L.: The iso standard for dialogue act annotation. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 549–558 (2020)

Bunt, H., Prasad, R.: Iso dr-core (iso 24617-8): Core concepts for the annotation of discourse relations. In: Proceedings 12th Joint ACL-ISO Workshop on Interop-erable Semantic Annotation (ISA-12). pp. 45–54 (2016)

Crible, L.: Identifying and describing discourse markers in spoken corpora. annotation protocol v.8. Tech. rep. (2014)

Crible, L., Degand, L.: Reliability vs. granularity in discourse annotation: What is the trade-off? Corpus Linguistics and Linguistic Theory 15(1), 71–99 (2019). https://doi.org/doi:10.1515/cllt-2016-0046, https://doi.org/10.1515/cllt- 2016-0046

Crible, L., Zufferey, S.: Using a unified taxonomy to annotate discourse markers in speech and writing (04 2015)

Das, D.: Signalling of coherence relations in discourse. Ph.D. thesis, Arts & Social Sciences: Department of Linguistics (2014)

Fraser, B.: Pragmatic markers. Pragmatics 6, 167–190 (1996)

González, M.: Pragmatic markers and discourse coherence relations inies 7(1), https://doi.org/10.1177/1461445605048767

english and catalan oral narrative. Discourse Stud- 53–86 (2005). https://doi.org/10.1177/1461445605048767,

Knott, A., Dale, R.: Using linguistic phenomena to motivate a set of rhetorical relations. Human Communication

References

Mann, W., Thompson, S.: Rethorical structure theory: Toward a functional theory of text organization. Text 8, 243–281 (01 1988).
https://doi.org/10.1515/text.1.1988.8.3.243

Maschler, Y., Schiffrin, D.: Discourse markers: Language, meaning, and context. The handbook of discourse analysis 1, 189–221 (2015)

Mendes,A.,delRíoGayo,I.,Stede,M.,Dombek,F.:Alexiconofdiscoursemarkers for portuguese - ldm-pt. In: LREC (2018)

Prasad, R., Bunt, H.: Semantic relations in discourse: The current state of ISO 24617-8. In: Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11). Association for Computational Linguistics, Lon- don, UK (Apr 2015), https://aclanthology.org/W15-0210

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Web- ber, B.: The Penn Discourse TreeBank 2.0. In: Proceedings of the Sixth Inter- national Conference on Language Resources and Evaluation (LREC’08). Euro- pean Language Resources Association (ELRA), Marrakech, Morocco (May 2008), http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf

Sanders, T., Spooren, W., Noordman, L.: Toward a taxonomy of coherence relations. Discourse Processes - DISCOURSE PROCESS 15, 1–35 (01 1992). https://doi.org/10.1080/01638539209544800

Sanders, T.J., Demberg, V., Hoek, J., Scholman, M.C., Asr, F.T., Zufferey, S., Evers-Vermeul, J.: Unifying dimensions in coherence relations: How various annota- tion frameworks are related. Corpus Linguistics and Linguistic Theory 17(1), 1–71 (2021). https://doi.org/doi:10.1515/cllt-2016-0078, https://doi.org/10.1515/cllt- 2016-0078

Schiffrin, D.: Discourse Markers. Cambridge University Press, Cambridge (1987).
https://doi.org/http://dx.doi.org/10.1017/CBO9780511611841

Silvano, M.d.P.M.: Temporal and rhetorical relations: the semantics of sentences with adverbial subordination in European Portuguese. Ph.D. thesis (2010)

Stede, M., Scheffler, T., Mendes, A.: Connective-lex: A web-based multilingual lexical resource for connectives. Discours (2019)

Taboada, M.: Discourse markers as signals (or not) of rhetorical relations. Journal of Pragmatics 38(4), 567–592 (2006).
https://doi.org/https://doi.org/10.1016/j.pragma.2005.09.010,

Zeyrek, D., Mendes, A., Kurfalı, M.: Multilingual extension of PDTB-style an- notation: The case of TED multilingual discourse bank. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC

Thanks!

ANY QUESTIONS?

Workshop Discourse studies and linguistic data science: Addressing challenges in interoperability, multilinguality and linguistic data processing DiSLiDaS 2022

ISO-DR-core plugs into ISO-dialogue acts for a crosslinguistic taxonomy of discourse markers

Purificação Silvano & Mariana Damova

University of Porto & Mozajka Ltd
msilvano@letras.up.pt & mariana.damova@mozajka.co

Nexus Linguarum

U.PORTO
FLUP FACULDADE DE LETRAS UNIVERSIDADE DO PORTO

MOZAIKA

2. Extra Resources

Background and Motivation

➤ Discourse Markers - Definition

- Schriffin (1987): "I operationally define markers as sequentially dependent elements which bracket units of talk." (includes connectives (*because*, *but*) and pragmatic particles (*you know*, *I mean*).

- Crible & Degaland (2019): "However, the functions of DMs go much further than this "bracketing" role.

Background and Motivation

- **Discourse Markers – Multifunctionalty**

  - Schriffin (1987): DM with roles in different dimensions: ideational structure, the exchange structure the information state or the participation framework.

  - Hovy (1995): DM convey semantic information, interpersonal purpose, and rhetorical relation.

  - Petukhova & Bunt (2009): DM can have “multiple meanings simultaneously which are related to the multiple purposes that an utterance may have in communication”

Background and Motivation

➤ Discourse markers - taxonomies

- ✓ Petukhova & Bunt (2009)
  - ○ Empirically-based and formal approach of the semantic functions of discourse markers in dialogue capable of capturing their multifunctional nature
  - ○ Semantic framework of Dynamic Interpretation Theory (Bunt, 2000): multilayered and multidimensional taxonomy with a set of communicative functions.

⬇

- ✓ Semantic annotation framework (SemAF) — Part 2: Dialogue acts - ISO 24617-2 (2010; 2020): interoperable dialogue act annotation framework

Background and Motivation

- **Taxonomies**

✓ Petukhova & Bunt (2009)

✓ Semantic annotation framework (SemAF) — Part 2: Dialogue acts - ISO 24617-2 (2010; 2020): interoperable dialogue act annotation framework

○ a wide-ranging metamodel for the annotation of dialogue acts that includes dimensions, communicative functions and qualifiers and plug-in to ISO 24617-8

although a interoperable, it doesn’t target specifically

Background and Motivation

- **Discourse markers - taxonomies**

✓ Prasad & Bunt (2015), Bunt & Prasad (2016), ISO 24617-8 (2016)

○ Interoperable set of core low-level semantic discourse relations according to the meaning of the relation’s arguments.

○ “a future part of ISO 24617 is envisaged that will complement this document by providing a complete interoperable annotation scheme for DRels, while also addressing the multilingual dimension of the standard”

Background and Motivation

- **Discourse markers - taxonomies**

✓ Crible (2017); Crible & Degand (2019)

○ Annotation taxonomy of DM in spoken language featuring two independent layers of semantic-pragmatic information, domains and functions

○ Four domains (González, 2005) - ideational, rhetorical, sequential or interpersonal – with 15 functions – addition, contrast,.. (some based on Prasad et al., 2007)

○ Tried out in different languages (French, English, Polish, Spanish) and modalities (spoken, written,