COSÌ: An Online Repertoire of Italian Discourse Markers

Miriam Voghera, Andrea Sansò, Iolanda Alfano, Alfonsina Buoniconto, Violetta Cataldo, Giovanni Di Paola, Pasquale Esposito, Flavio Pisciotta, Carmela Sammarco, Luisa Troncone

Abstract

From the perspective of the current landscape of linguistic resources, discourse markers (henceforth DMs), i.e. procedural elements such as *you know* and *I mean* that play a central role in structuring spoken interaction (Schiffrin 1987, Jucker & Ziv 1998, Fedriani & Sansò 2017, Sansò 2020, among many others) are markedly underrepresented. On the one hand, there are annotation protocols that aim to be general and not language-specific, but they raise issues regarding annotation choices and the exclusion of certain elements whose boundaries are difficult to define. One such example is the annotation protocol designed by Bolly et al. (2015; see also Crible & Zufferey 2015; Crible & Cuenca 2017), which sets very clear boundaries that exclude from annotation any elements falling into the categories of fillers, interjections, response signals, epistemic parentheticals, general extenders, tag questions, and editing terms. On the other hand, there exists only one digital repertoire of DMs, the *Diccionario de Partículas Discursivas del Español* (DPDE; Briz et al. 2008), designed as a dictionary with real examples from written and spoken Spanish, but entirely self-contained, with no links to other resources such as corpora, and with only limited access to examples in context. As for Italian, no such resource is currently available.

The aim of this paper is to present the initial stages in the development of COSÌ (*Catalogo On-line dei Segnali discorsivi Italiani*), an online repertoire of Italian DMs, conceived in an innovative way as a resource that is both linked to a spoken corpus and based on a well-tested and refined annotation protocol. What sets this resource apart is its tight integration with the *KiParla* corpus (Mauri et al. 2019). This integration is bidirectional: (i) the selection of potential candidates for discourse marker (DM) annotation is based on frequency lists derived from the KiParla corpus; and (ii) the annotation of DMs is implemented as an additional layer within KiParla, following a functional classification inspired by widely recognized categories in the literature (e.g., reformulation markers, turn-taking devices, etc.). Each entry in the repertoire includes (i) usage frequency, (ii) functional descriptions, and (iii) sociolinguistic distributions. The repertoire also provides direct access to corpus examples in context—with the possibility of listening to the audio files and autonomously analyzing prosodic features.

The annotation of DMs poses specific methodological challenges: their boundaries are often fuzzy, their functions highly context-dependent, and cases of multifunctionality are ubiquitous. Rather than imposing rigid predefined categories, the project adopts an intermediate level of annotation—rich enough to capture functional nuances and emergent patterns, yet modest enough to avoid overinterpretation and to leave space for inductive, data-driven analysis. This design aims to support both qualitative exploration and quantitative analysis, while preserving the variability and interactional embeddedness of these phenomena.

More specifically, we report on two annotation experiments designed to test the replicability of our functional tag set. The process unfolded in two phases: annotators first worked independently, without prior discussion, applying an initial version of the annotation scheme to a shared dataset of examples of a specific DM. After reviewing disagreements and identifying recurring issues, the annotation protocol was revised to include a set of strict criteria (e.g., semantic-syntactic independence and syntactic optionality of the discourse marker), along with a step-by-step checklist to help annotators distinguish genuine DM from non-DM uses of the same items.

The results of this first annotation campaign highlight the need for extensive training and discussion, even among expert annotators, and point to the importance of a more prescriptive and fine-grained operational definition of DMs than originally provided. By anchoring the repertoire in a spoken corpus and balancing analytic precision with openness to data-driven insights, the project fills a significant gap in resources for Italian and offers a replicable model for similar work in other languages.

References

Bolly, C. T., L. Crible, L. Degand & D. Uygur-Distexhe. 2015. MDMA. Un modèle pour l'identification et l'annotation des marqueurs discursifs "potentiels" en context. *Discours* 16: 3-32.

Briz, A., S. Pons Bordería, & J. Portolés (eds.). 2008. *Diccionario de partículas discursivas del español*. On-line, www.dpde.es.

Crible, L. & M. J. Cuenca. 2017. Discourse markers in speech: Characteristics and challenges for corpus annotation. *Dialogue and Discourse* 8 (2): 149-166. https://doi.org/10.5087/dad

Crible, L. & S. Zufferey. 2015. Using a unified taxonomy to annotate discourse markers in speech and writing. In H. Bunt (ed.), *Proceedings of the 11th Joint ACL - ISO Workshop on Interoperable Semantic Annotation (isa-11)*, 14.22.

Fedriani, C. & A. Sansò. 2017. Introduction. Pragmatic Markers, Discourse Markers and Modal Particles: What do we know and where do we go from here? In C. Fedriani & A. Sansò (eds.), *Pragmatic markers, discourse markers and modal particles: New perspectives*, 1-33. Amsterdam: John Benjamins.

Jucker, A. H. & Y. Ziv. 1998. Discourse Markers: Introduction. In A. H. Jucker & Y. Ziv (eds.), *Discourse Markers*, 1-12. Amsterdam: John Benjamins.

Mauri, C., S. Ballarè, E. Goria, M. Cerruti & F. Suriano. 2019. KIParla corpus: a new resource for spoken Italian". In R. Bernardi, R. Navigli & G. Semeraro (eds.), *Proceedings of the 6th Italian Conference on Computational Linguistics CLiC-it*.

Sansò, A. 2020. *I segnali discorsivi*. Roma: Carocci.

Schiffrin, D. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.