Segmenting a French Meeting Corpus into Elementary Discourse Units

Laurent Prévot
Aix Marseille Univ. & CNRS, LPL
CNRS & MEAE, CEFC
laurent.prevot@univ-amu.fr

Roxane Bertrand
CNRS & Aix Marseille Univ., LPL
Aix-en-Provence, France
roxane.bertrand@univ-amu.fr

Julie Hunter
LINAGORA Labs
Toulouse, France
jhunter@linagora.com

Abstract

Despite growing interest in discourse-related tasks, the limited quantity and diversity of discourse-annotated data remain a major issue. Existing resources are largely based on written corpora, while spoken conversational genres are underrepresented. Although discourse segmentation into elementary discourse units (EDUs) is considered to be nearly solved for canonical written texts, conversational spontaneous speech transcripts present different challenges. In this paper, we introduce a large French corpus of segmented meeting dialogues, including 20 hours of manually transcribed and discourse-annotated conversations, and 80 hours of automatically transcribed and segmented data. We describe our annotation campaign, discuss inter-annotator agreement and segmentation guidelines, and present results from fine-tuning a model for EDU segmentation on this resource.

1 Introduction

Discourse processing is gaining increasing attention as performance levels have reached a threshold of practical usability. However, the field still faces significant challenges, due to the limited quantity and diversity of discourse-annotated data. Most existing discourse resources are based on written corpora, while spoken conversational genres remain underrepresented. Moreover, the largest available resources for spoken discourse are typically limited to English telephone conversations between two participants, despite the broader diversity of real-world conversational settings. In this paper, we introduce a new, sizable French corpus composed of meetings involving 3 to 4 participants.

The basic units of discourse processing, known as *elementary discourse units* (EDUs), are often approximated using punctuation-based sentence boundaries in written texts, or pauses and speaker changes in spoken dialogue. However, such proxies are imperfect, and *discourse segmentation* is widely recognized as a necessary step for accurate discourse analysis. While state-of-the-art results suggest that discourse segmentation is nearly solved for written texts—thanks to the presence of reliable punctuation—this is not the case for spontaneous speech. To address this gap, we present the creation of a large, discourse-segmented corpus of French meetings, comprising approximately 20 hours of manually transcribed and segmented conversations, along with 80 hours of automatically transcribed and segmented data.

The paper begins with a survey of existing discourse corpora (Section 2) and related work on discourse segmentation (Section 3), before introducing the French meeting corpus (Section 4). We then describe the discourse segmentation campaign (Section 5), followed by a proposal for a baseline model, fine-tuned on our annotations and applied to the full dataset (Section 6).

2 Discourse Segmented Conversational Corpora

The AMI corpus (Carletta et al., 2005) and the ICSI corpus (Janin et al., 2003; McCowan et al., 2005) are among the most well-known resources for meeting conversations. AMI includes 137 scenario-driven meetings, each lasting between 15 and 45 minutes, totaling approximately 65 hours of conversation. In each meeting, four participants assume specific roles and interact over a sequence of four sessions. While the language is spontaneous, the scripted design results in conversational styles that are arguably cleaner and more structured than those found in real-life meetings, leading to greater homogeneity in both content and vocabulary.

The ICSI corpus consists of natural meetings,

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 183–191
Aug 25–27, 2025.

each about one hour long, totaling roughly 72 hours of recordings. These meetings typically involve six participants discussing technical topics related to natural language processing. Since ICSI features real meetings between collaborators actively working on projects at the time of recording, the interactions are more naturalistic than those in AMI. However, the conversations include a high concentration of technical vocabulary, specialized subject matter, and significant shared knowledge. These factors may complicate interpretation for models lacking this contextual background.

The ELITR corpus (Nedoluzhko et al., 2022) contains transcripts of 113 technical project meetings in English and 53 in Czech, totaling over 160 hours of content. The transcriptions include turn-level segmentation, with turns further punctuated using standard written punctuation marks.

Additionally, several smaller conversational corpora are available in French (for a recent list, see Hunter et al., 2023). One such example is CID (Corpus of Interactional Data; Blache et al., 2017), which includes eight one-hour dialogues between friends, annotated with elementary discourse segmentation (Prévot et al., 2021). However, with only eight hours of recording, this corpus remains relatively small and is not specifically focused on meetings.

3 Discourse Segmentation

Discourse segmentation involves breaking down written texts or speech transcripts into units that reflect the communicative intentions of the participants. These units aim to provide a more accurate representation of discourse structure than traditional notions such as *'sentences'* or *'speech activity chunks'* (e.g., interpausal units). We use the term *elementary discourse unit (EDU)* from Asher and Lascarides (2003) to refer to minimal communicative units, each corresponding to clause-level content that conveys a single fact or event.

For a long time, segmentation into EDUs remained a challenging task, requiring significant human annotation effort to build discourse-segmented corpora. The development of LLM-based approaches has opened in a new era, in which these tasks can now be tackled with high accuracy, particularly for written genres and highly resourced languages.

Large Language Models (LLMs) have achieved strong results in discourse segmentation (Zeldes et al., 2019, 2021; Braud et al., 2023). The work carried out within these shared tasks has equipped the community with robust tools and frameworks for EDU segmentation. However, as noted by Braud et al. (2023), performance tends to decline for languages other than English, even in written genres, particularly when gold-standard sentence boundaries are unavailable, due to the imperfections of automatic sentence segmenters (Braud et al., 2017). While recent work such as Minixhofer et al. (2023) and Frohmann et al. (2024) is less reliant on punctuation, it remains anchored in textual domains.

A recent trend in discourse segmentation leverages sequential models over contextual embeddings (Wang et al., 2018; Muller et al., 2019). Specifically for spontaneous speech, Gravellier et al. (2021) proposed a weak supervision approach (Ratner et al., 2017), where manually crafted heuristic rules, including some based on a model trained on written data (Muller et al., 2019), were used to annotate the corpus. This noisy supervision was then used to fine-tune BERT (Devlin et al., 2019). In Prevot et al. (2023), a larger set of manual annotations enabled a direct comparison between fine-tuning with a sizable training dataset and weak supervision. Metheniti et al. (2023a) introduced improvements over Muller et al. (2019), achieving new state-of-the-art results across multiple languages. Building on this line of work, Prevot and Wang (2024); Prevot et al. (2025) demonstrated that fine-tuning ROBERTA provides a promising general approach to discourse segmentation, even for lower-resource languages.

Working with conversational speech transcripts, rather than written text, introduces several challenges. First, there is no ground-truth punctuation in such data; instead, it is replaced by pauses and, in multimodal setups, prosody (which is not the case in this paper, but see (Gravellier et al., 2021; Prevot and Wang, 2024) for similar experiments involving prosody). The specific flow of spontaneous speech introduces another difficulty, as disfluencies are frequent and often disrupt canonical syntactic structures.

Moreover, the conversational nature of the data introduces two additional challenges. First, turn alternation can significantly disrupt the sequences. Second, discourse units exhibit much greater variability in content compared to written documents. More precisely, they include explanatory, narrative,

and other monologic sequences, as well as shorter interactional units corresponding to conversational feedback and short answers.

Regarding major dialogic phenomena, whenever a *backchannel* corresponds to a full dialogue act, it is segmented accordingly. It is important to note that segmentation is performed independently for each participant, although annotators are allowed to see the full multilogue context.

In the case of *hesitations*, we distinguish between completely abandoned false starts—which cannot be semantically or syntactically related to what follows (e.g., due to an inconsistent subject)—and hesitations that are disfluent yet coherent preparations for the upcoming content. In the former case, the abandoned units are segmented as separate discourse units; in the latter, they are grouped with the following segment.

Although laughter is a crucial phenomenon in dialogue (Ginzburg et al., 2020), it was not considered in our segmentation work.

4 The Corpus

We provide elementary discourse segmentation for the SUMM-RE corpus (Hunter et al., 2024).[1] This corpus consists of approximately 100 sessions made of 3 x 20' meeting, with 2 to 4 participants, focused on event planning. Each group takes part in three sessions, each associated with a different task: (i) discussing ideas for the event (more presentation- and monologue-oriented), (ii) deciding what to do at the event (dialogic discussions involving divergent opinions), and (iii) planning the practical organization of the event (including task delegation). Most sessions were recorded face-to-face using head-mounted microphones, with a few additional sessions conducted via Zoom.

The entire SUMM-RE corpus has been automatically transcribed (Yamasaki et al., 2023). The full `dev` and `test` sets— corresponding to 73 meetings and approximately 24 hours of recordings—have been manually corrected and annotated. While this may not be considered a large dataset, to the best of our knowledge, it is the largest annotated corpus available for French and among the largest across all languages. We computed some basic figures of the manually annotated part of the corpus, the dialogue hosted in average 534 EDUs and mean length of an EDUs is 8 ± 6 (exhibiting a huge variation due to the diversity of the speech acts conveyed by these discourse units). See Figures 3 and 4 in the Appendix for details.

[1] https://huggingface.co/datasets/linagora/SUMM-RE

5 Discourse Segmentation Campaign

Four naive annotators[2] were recruited to manually segment the `dev` and `test` portions of the corpus. The process began with a discussion of segmentation guidelines—adapted from earlier work on written French (Muller et al., 2012) and previously applied to similar datasets (Prévot et al., 2021)—followed by training sessions during which all debatable decisions made by the annotators were discussed and potentially revised. We held three of such sessions, using both an external corpus (Blache et al., 2009) and a few meetings from our own dataset. After each session, annotators conducted a short segmentation exercise, and inter-annotator agreement was evaluated. Once convergence was sufficient, as measured by the inter-annotator κ score, we proceeded with the full annotation task. While discourse segmentation involves many straightforward decisions, some cases can be quite challenging—particularly in the context of spontaneous conversational speech.

The segmentation was then performed using Praat (Boersma, 2002) to avoid requiring specialized tool training and to ensure precise timing of each participant's contributions. Annotation was conducted speaker by speaker, with the other participants' contributions used as contextual information. Thus, for a conversation involving four participants, a separate segmentation file was created for each speaker, and the session was analyzed four times.

5.1 A Two-Step Approach

The annotation process was carried out in two steps. First, a small subset of the corpus was segmented manually by the naive coders. This subset was then used to train an initial rough segmentation model. We applied this model to the entire corpus to produce pre-segmented files, which the coders subsequently edited. As previously mentioned, segmentation involves a majority of easy cases that the model was able to learn quickly. The main benefit of this approach was a significant reduction in annotation time: Annotators only needed to edit an existing Praat file rather than create all the time boundaries from scratch—a notably tedious task.

[2] Undergraduate students from the School of Humanities.

Even a rough baseline pre-segmentation proved extremely helpful in this context.

Such a two-step method raises the concern that the resulting dataset might be overly biased by the pre-segmentation. To address this, we conducted two checks. First, we manually compared the pre-segmentation and the final manual segmentation on a sample of sessions and found that the annotators made effective use of the pre-segmentation. For instance, spontaneous dialogue often includes short feedback utterances, which are preceded and followed by long speaker pauses; while such segments require multiple operations to annotate from scratch, they were consistently handled correctly in the pre-segmentation.

Second, we performed inter-coder agreement analysis on both files annotated from scratch and those annotated with pre-segmentation, and observed no significant differences between the two. This does not mean that the dataset is entirely free of bias introduced by pre-segmentation, but if any bias exists, it appears to be smaller than the variability introduced by inter-coder differences, as shown in Figure 1. Overall, semi-naive coder reached mean κ-scores within the 0.85-0.89 range, while a few expert annotations conduced by the authors of the paper have shown higher reliability, with κ-scores around 0.9.

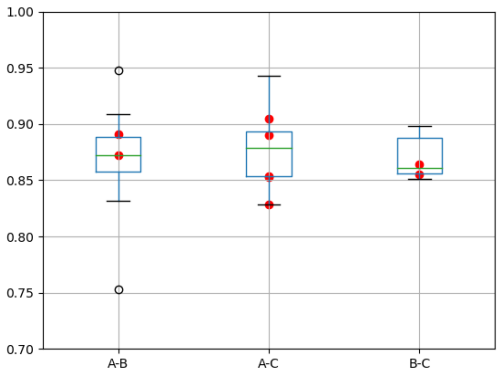

Figure 1: F-scores between human coders. Red dots correspond to agreement on files that were not pre-segmented.

5.2 Segmentation Guidelines

We adopted a discourse approach that is designed to be simple enough for semi-naive coders to apply after some training. While we acknowledge the multidimensional nature of discourse phenomena, our approach focuses on a single layer of discourse units. Although multiple overlapping layers are analytically justified (Petukhova et al., 2011; Hu and Degand, 2023), we believe that both conversation participants and semi-naive coders can intuitively identify discourse units as conversational actions, without explicitly attending to these complex overlapping layers. To summarize, our definition of EDU is proposed below, in (1).

Our guidelines use both holistic interpretation of discourse function within the discourse flow as understood by the annotator and surface cues such as discourse markers. This operationalization combines both top-down and bottom-up perspectives. This reflects the intermediate position of EDUs, which function as the largest units within the domain of grammar and language, and the smallest units within discourse structure. The complete segmentation guidelines are provided as Supplementary Material.

(1) **Elementary Discourse Unit (EDU)** In a speech transcript, an EDU is defined as a span of contiguous tokens whose combined meaning corresponds to a semantic proposition (describing an event, fact, or state of affairs), or a single speech act, such as asking a question, providing an answer, or offering conversational feedback. All preparatory disfluent material should be included within the unit.

A unit is only labeled as an *abandoned discourse unit* if the introduced material is discarded before the beginning of a different discourse unit.

6 Automatic Discourse Segmentation

The primary focus of our experiments is on fine-tuning large language models (LLMs), specifically `xlm-roberta-large` (Conneau et al., 2019). Our models follow a sequence-to-sequence architecture, fine-tuned using the methodology proposed by Metheniti et al. (2023b) within the JIANT framework (Pruksachatkun et al., 2020).[3]

The task is defined within the DISRPT framework. Data is encoded in the CONLL format, with a simple binary label indicating whether a token marks the beginning of a discourse unit (see Table 2). Evaluation is based on F-score, precision, and recall, computed exclusively for discourse boundaries (true negatives are excluded from the

[3]See https://github.com/phimit/jiant-discut

score).

We used XLM-RoBERTa (Conneau et al., 2019) with a learning rate of 10^{-5}, a batch size of 1, gradient accumulation of 4, and a maximum of 30 epochs with a patience of 10, based on performance on the development set. Results by training set size are reported in Table 2.

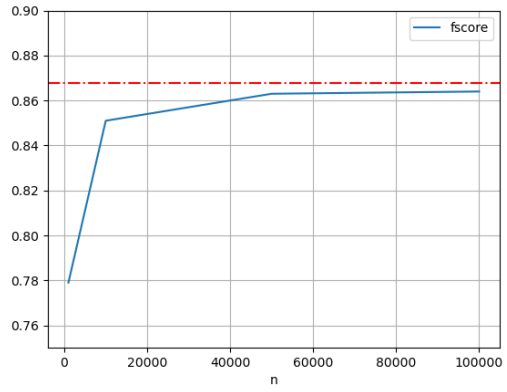

Figure 2: F-scores according to training data size. In red, the mean value for human annotations.

The results in Figure 2 indicate that the model's performance plateaus relatively quickly. Once the naive coders were trained, we estimated that segmenting 1,000 tokens into EDUs would take approximately half a day per person, 10,000 tokens about a week, 50,000 tokens around a month, and 100,000 tokens roughly two months. Notably, this performance plateau seems to correspond to the maximal theoretical level achievable, as it aligns with the inter-coder agreement levels observed in our study. It appears that we performed more manual segmentation than was strictly necessary to achieve optimal model performance. Nevertheless, this extensive gold-standard dataset will be valuable for future research and could contribute to the development of more robust discourse segmentation models.

7 Conclusion

In this paper, we introduced a manual segmentation of a large meeting corpus and presented a straightforward fine-tuning approach to build a segmentation model. The model, along with its application to the full dataset, will be released alongside the code repository.

Our findings indicate that fine-tuning for this task is highly effective, requiring relatively modest amounts of annotated data to achieve strong performance. Notably, the model's performance plateaued after a certain point, aligning with the inter-coder agreement levels observed in our study.

Nevertheless, the extensive gold-standard annotations we produced remain valuable for future studies. From a linguistic perspective, this substantial (20-hour) manually discourse-segmented corpus offers a rich resource for exploring the discourse-prosody interface in multiparty conversations.

For future work, we are interested in assessing the extent to which discourse segmentation can be effectively applied to the ASR versions of the corpus. While an F-score of approximately 0.86 makes the method more promising than typical proxies, it still falls short of the scores typically achieved on written documents for this task. One potential avenue we would like to pursue is collecting more segmentations from expert (rather than naive) annotators. Given the efficiency of fine-tuning observed in our study, fine-tuning on a small amount of expert annotations or using them in a many-shot approach (Agarwal et al., 2024) could be a promising strategy.

Acknowledgments

We would like to thank our discourse annototators: Eliane Bailly, Océane Granier, Manon Méaume and Lyne Rahabi. We gratefully acknowledge support from the ANR-funded project, SUMM-RE (ANR-20-CE23-0017).

References

Rishabh Agarwal, Avi Singh, Lei Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. 2024. Many-shot in-context learning. *Advances in Neural Information Processing Systems*, 37:76930–76966.

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

Philippe Blache, Roxane Bertrand, and Gaëlle Ferré. 2009. Creating and exploiting multimodal annotated corpora: the toma project. *Multimodal corpora*, pages 38–53.

Philippe Blache, Roxane Bertrand, Gaëlle Ferré, Berthille Pallaud, Laurent Prévot, and Stéphane Rauzy. 2017. The corpus of interactional data: A large multimodal annotated resource. *Handbook of linguistic annotation*, pages 1323–1356.

P. Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017. Does syntax help discourse segmentation? not so much. In *Conference on Empirical Methods in Natural Language Processing*, pages 2432–2442.

Chloé Braud, 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 *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11908–11941, Miami, Florida, USA. Association for Computational Linguistics.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa: a journal of general linguistics (2021-...)*, 5(1).

Lila Gravellier, Julie Hunter, Philippe Muller, Thomas Pellegrini, and Isabelle Ferrané. 2021. Weakly supervised discourse segmentation for multiparty oral conversations. In *Proceedings of EMNLP 2021*.

Junfei Hu and Liesbeth Degand. 2023. The conversational discourse unit: Identification and its role in conversational turn-taking management. *Dialogue & Discourse*, 14(2):83–112.

Julie Hunter, Jérôme Louradour, Virgile Rennard, Ismaïl Harrando, Guokan Shang, and Jean-Pierre Lorré. 2023. The claire french dialogue dataset. *arXiv preprint arXiv:2311.16840*.

Julie Hunter, Hiroyoshi Yamasaki, Océane Granier, Jérôme Louradour, Roxane Bertrand, Kate Thompson, and Laurent Prévot. 2024. MEETING: A corpus of French meeting-style conversations. In *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, pages 508–529, Toulouse, France. ATALA and AFPC.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The ICSI meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP'03)*, volume 1, pages I–I. IEEE.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Wilfried Post, Dennis Reidsma, and P. Wellner. 2005. The AMI meeting corpus. *International Conference on Methods and Techniques in Behavioral Research*.

Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 2023a. DisCut and DiscReT: MELODI at DISRPT 2023. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42, Toronto, Canada. The Association for Computational Linguistics.

Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 2023b. Discut and discret: Melodi at disrpt 2023. In *3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 29–42. ACL: Association for Computational Linguistics.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada. Association for Computational Linguistics.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. Tony: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124. Association for Computational Linguistics.

Philippe Muller, Marianne Vergez-Couret, Laurent Prévot, Nicholas Asher, Benamara Farah, Myriam Bras, Anne Le Draoulec, and Laure Vieu. 2012. Manuel d'annotation en relations de discours du projet annodis. Technical Report 21, CLLE-ERS, Toulouse University.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR

Minuting Corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC-2022)*, Marseille, France. European Language Resources Association (ELRA). In print.

Volha Petukhova, Laurent Prévot, and Harry Bunt. 2011. Multi-level discourse relations between dialogue units. In *Proceedings 6th joint ACL-ISO workshop on interoperable semantic annotation (ISA-6), Oxford*, pages 18–27.

Laurent Prévot, Roxane Bertrand, and Stéphane Rauzy. 2021. Investigating disfluencies contribution to discourse-prosody mismatches in french conversations. In *The 10th Workshop on Disfluency in Spontaneous Speech*.

Laurent Prevot, Julie Hunter, and Philippe Muller. 2023. Comparing methods for segmenting elementary discourse units in a French conversational corpus. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 436–446, Tórshavn, Faroe Islands. University of Tartu Library.

Laurent Prevot, Philippe Muller, Shu-Chuan Tseng, and Sheng-Fu Wang. 2025. Llm-based elementary discourse units segmentation for spontaneous speech in french, taiwan mandarin, and taiwan southern min. *International Journal of Computational Linguistics and Chinese Language Processing*, 30(2).

Laurent Prevot and Sheng-Fu Wang. 2024. Experimenting with discourse segmentation of taiwan southern min spontaneous speech. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 50–63.

Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R Bowman. 2020. jiant: A software toolkit for research on general-purpose text understanding models. *arXiv preprint arXiv:2003.02249*.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Hiro Yamasaki, Jérôme Louradour, Julie Hunter, and Laurent Prévot. 2023. Transcribing and aligning conversational speech: A hybrid pipeline applied to french conversations. In *2023 IEEE Automatic Speech Recognition and Understanding*.

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 2019. The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene, editors. 2021. *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*. Association for Computational Linguistics, Punta Cana, Dominican Republic.

A Appendix

(2) Eventualities

a. [on y va avec des copains]$_{du}$ [on avait pris le ferry en Normandie]$_{du}$
[we are going there with friends]$_{du}$ [we took the ferry in Normandy]$_{du}$
[puisque j'avais un frère qui était en Normandie]$_{du}$ [on traverse]$_{du}$
[since I had a brother that was in Normandy]$_{du}$ [we cross]$_{du}$
[on avait passé une nuit épouvantable sur le ferry]$_{du}$
[we spent a terrible night on the ferry]$_{du}$

b. [j'ai eu plusieurs conflits avec des animateurs pas assez sérieux]$_{du}$
[I had several conflicts with group leaders that were not serious enough]$_{du}$

c. [et y en a un qui s'était pris un banc de pierre]$_{du}$
[and there was one that hit a stone bench]$_{du}$

(3) Speech Act / Clear Communicative Function

a. A: [Tu vois où c'est?]$_{du}$ B: [oui]$_{du}$
A: [You know where it is?]$_{du}$ B: [Yes]$_{du}$

b. A: [Je ne voulais pas les déranger]$_{du}$ B: [oui bien sûr]$_{du}$
A: [I did not want to disturb them]$_{du}$ B: [Yes of course]$_{du}$

(4) Discourse Markers inducing segmentation

a. [on a appelé euh des les parents d'amis]$_{du}$ [mais pas d'amis de notre âge d'amis de mes parents]$_{du}$
[we called um some friend's parents]$_{du}$ [but not friends of our age friends or my parents]$_{du}$

b. [donc on était à Montréal en fait]$_{du}$ [et après le congrès on est parti en Gaspésie]$_{du}$
[so we were in Montreal in fact]$_{du}$ [and after the conference we left to Gaspésie]$_{du}$

nb train tokens	prec	recall	fscore
1000	0.869	0.706	0.779
10000	0.880	0.826	0.851
50000	0.876	0.850	0.863
100000	0.900	0.831	0.864

Table 1: Results for Discourse Segmentation for different amount of training gold data.

id	token	connll	Discourse Boundary
1	ouais	_ _ _ _ _ _ _	BeginSeg=Yes
2	#	_ _ _ _ _ _ _	_
3	on	_ _ _ _ _ _ _	BeginSeg=Yes
4	dirait	_ _ _ _ _ _ _	_
5	des	_ _ _ _ _ _ _	_
6	enfants	_ _ _ _ _ _ _	_
7	#	_ _ _ _ _ _ _	_
8	hein	_ _ _ _ _ _ _	_
9	#	_ _ _ _ _ _ _	_
10	mais	_ _ _ _ _ _ _	BeginSeg=Yes
11	les	_ _ _ _ _ _ _	_
12	enfants	_ _ _ _ _ _ _	_

Table 2: Illustration of the data format for the discourse segmentation task.

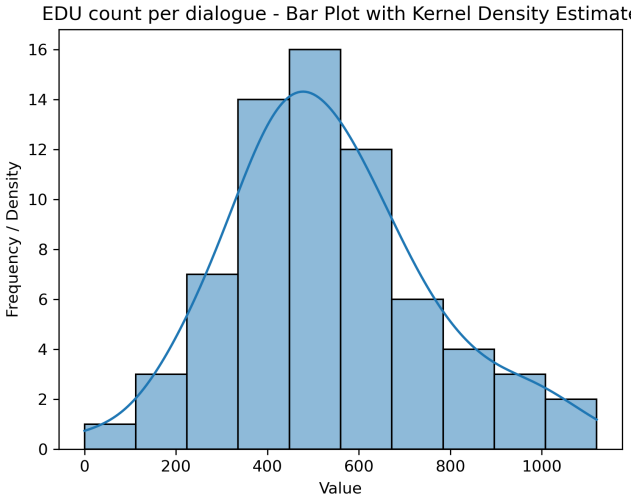

Figure 3: EDU counts per dialogue

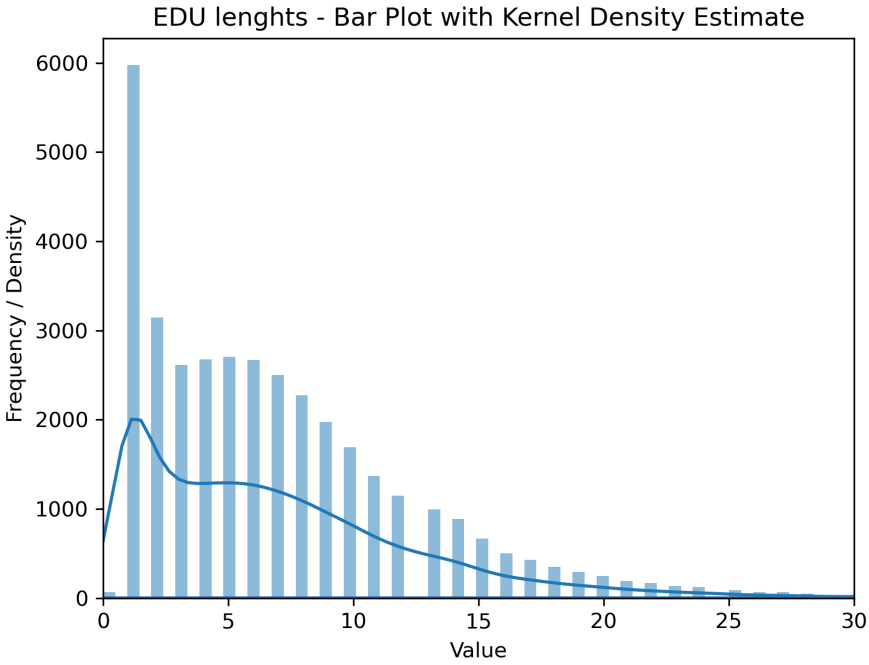

Figure 4: EDU lengths distribution