Dialogue & Discourse 16(2) (2025) 35–73 doi: 10.5210/dad.2025.202

A few shades of supervision for discourse segmentation: experiments on a French conversational corpus

Laurent Prévot LAURENT.PREVOT@UNIV-AMU.FR
CNRS & MEAE, CEFC, Taipei, Taiwan
Aix Marseille Univ & CNRS, LPL, Aix-en-Provence, France

Philippe Muller PHILIPPE.MULLER@IRIT.FR
IRIT, Université de Toulouse & ANITI, Toulouse, France

Editor: Massimo Poesio

Submitted 03/2025; Accepted 11/2025; Published online 12/2025

Abstract

Elementary Discourse Units (EDUs) constitutes the interface between language grammar and language use. One the one hand, they result from compositional semantic processes that combines individual word meanings into proposition-level representations. On the other hand, EDUs form the building blocks of most text, discourse, and dialogue frameworks. In written genres, where punctuation is available and reliable, segmenting EDUs is sometimes seen as a nearly solved problem, as least for high-resource languages. However, this is not the case for spontaneous speech transcripts. In this paper, we use a significant (8-hour) French corpus, manually segmented into EDUs, to evaluate several large language model (LLM)-based approaches for this task. We compare various fine-tuning strategies, including those relying on weakly supervised labels, in relation to the amount of "gold" manual annotations that can be available. We also experiment with in-context learning, where example instances are provided to condition a generative model (few-shots learning) or in a purely generative approach (zero-shot). Our findings indicate that classical fine-tuning is still the most effective approach, requiring only a reasonable amount of gold-annotated data to achieve the best performance in our experiments. Beyond traditional quantitative evaluation, we conducted a systematic qualitative analysis, identifying directions for further improvement. These include integrating prosodic considerations while handling pauses when they co-occur with disfluencies or complex discourse markers uses. Finally, we argue for the significance of this task and the resulting units, compared to acoustic and syntactic proxies, especially for quantitative linguistics focusing on spontaneous speech.

Keywords: Discourse Units, Conversation, Dialogue, LLM, Weak-Supervision

1. Introduction

Elementary Discourse Units (EDUs)[1] are both the maximal unit of traditional grammar (commonly referred to as the sentence) and the minimal unit of discourse analysis, that extends beyond the sen-

1. "Discourse Units" (DU) are often separated into "Elementary Discourse Units" (EDU) and Complex Discourse Units (CDU) that corresponds to sets of EDUs related by discourse relations (Asher and Lascarides, 2003). We focus in this paper on EDUs.

tence.[2] EDUs serve as the crucial articulation point between language structure and language use. They address the dual question: how do we form "ideas", or more precisely, semantic propositions, using words and grammar, and how do we use these "ideas" to communicate through discourse and dialogue?

(1) [et je cherchais ce cette communication]$_A$ [parcequ'elle était de folie]$_A$ [il en avait fait des photocopies]$_A$ [qu'il avait mises en pile dans un coin]$_A$ [quel barjot]$_B$ [et puis je les ai jamais retrouvé]$_A$ [tu les as jamais trouvé ouais]$_B$ [et voilà j'aurai bien aimé avoir ce truc de dingue]$_A$
[I was looking for this that talk]$_A$ [because it was just insane]$_A$ [he'd made some photocopies]$_A$ [and stacked them in a corner somewhere]$_A$ [what a weirdo]$_B$ [and then I never found them again]$_A$ [you never found them yeah]$_B$ [and that's it I wish I still had that crazy thing]$_A$[3]

Given their pivotal role, one might expect EDUs to be a central focus of many linguistic and natural language processing (NLP) studies. However, the opposite is true. Conversation, Dialogue or Communication studies tend to treat them as pre-existing units, while syntax loosely characterizes them as maximal projections. For written genres, this situation can be explained by the reliability of punctuation as a segmentation cue. However, for spontaneous speech—where no such surface cues exist—the neglect of this segmentation problem is puzzling. One possible explanation for this situation might lie in EDUs' position at the intersection of two linguistic domains and research communities. Analyzing discourse units requires insights from compositional semantics to understand how they are formed, but also from prosody and higher-level pragmatics related to speakers' intentions.

In natural language processing, early efforts (Polanyi and Scha, 1983; Passonneau and Litman, 1997) segmented text and speech flow (Hirschberg and Grosz, 1992), recognizing EDUs as potentially useful units for various applications (e.g., discourse summarization, cf Marcu, 2000). However, discourse processing and parsing were considered as niche tasks due to their complexity and the low performance of early systems. Large Language Models (LLMs) have changed this landscape. Discourse tasks gained attention in NLP conferences, and are supported by specialized events and workshops. Even so, "segmentation" remains an often-overlooked preliminary step, subordinate to downstream tasks like discourse parsing, argumentation structure analysis, or discourse summarization. See however, (Stede, 2012), who provides a detailed overview of segmentation, and the DISRPT initiative of Zeldes et al. (2019, 2021); Braud et al. (2023), reflecting a community interest in this area. Finally, most existing work focuses either on written genres or on highly interactional dialogues (e.g., task-oriented or chit-chat), which differ significantly from everyday conversation that combines various kind of conversational activities (e.g alternating storytelling with chit-chat sequences) of the kind we study here.

2. We will later discuss in detail the relationship between the notion of 'sentence' and discourse units. We use 'sentence' in this introduction as it is the most familiar term corresponding to both the syntactic maximal and discursive minimal unit.
3. In this example, EDUs are segmented into brackets and the A/B mentioned correspond to the speaker information.

2. Related work

Work on discourse, including topics such as discourse units, discourse markers, and topic changes, has focused on written genres, both in linguistics and NLP. This tendency reflects the documented "written bias" in these fields (Linell, 2004). In contrast, spontaneous conversational speech, while being a subject of interest within certain linguistic subfields, has been somehow overlooked by mainstream linguistics despite its critical importance for understanding language.

Discourse segmentation has however been addressed for both written and spoken data, including monologues and dialogues, though with differing lenses. Broadly, this research spans at least two levels of segmentation: (1) sentence-, utterance- or clause-like units and (2) paragraph- or topic-like units. The latter has been extensively studied in natural language processing for both written texts (Hearst, 1994) and spoken data (Passonneau and Litman, 1997). The former has received more interest from semanticists and discourse analysts as the basic unit of analysis, sometimes referred to explicitly as EDU (Polanyi and Scha, 1983; Stede, 2012; Asher and Lascarides, 2003). Relational approaches to discourse have used EDUs as foundational building blocks for constructing discourse structures.

We use example (1) to introduce the terminology for this section. We use *utterance* as the common sense notion of the result of performing one speech act. *Speech acts* follow Austin's traditional definition and are combination of a propositional content and an illocutionary force (e.g. asking, asserting,...) . They have to been generalized to *dialogue acts* (synonymous to *dialogue moves* or *communicative acts* that both can include various communicative functions such as communicative feedback, potentially related to other level of dialogue management (e.g. turn-taking).

2.1 Linguistics

There has been significant linguistic interest in discourse units in speech for a long time, beginning with work by Conversation Analysts (Sacks et al., 1974; Schegloff and Sacks, 1973) and later by Interactional Linguists (Ford and Thompson, 1996; Selting, 2000). Discourse analysts across various frameworks have also contributed, favoring functional over formal approaches and emphasizing corpus-based empirical methods (Sinclair and Coulthard, 1992; Brazil, 1995; Roulet et al., 2001). Scholars specializing in spoken language syntax have further contributed to the question (Pietrandrea and Kahane, 2019; Haselow, 2017). Degand and Simon (2005, 2009) already proposed to synthesize these proposals to define BASIC DISCOURSE UNITS. An overview of functional approaches to discourse in Romance languages is provided in (Pons Bordería, 2014). Finally, researchers exploring the formal semantics and pragmatics of dialogue, such as (Asher and Lascarides, 2003) and (Ginzburg, 2012), have also examined discourse units or dialogue moves in conversation.

2.1.1 INTERACTIONAL UNITS

Interactional Linguistics (Couper-Kuhlen and Selting, 2001), a development of conversation analysis focused on language resources, introduces *interactional units* (Ford and Thompson, 1996) as a crucial step for studying interactions. We consider these interactional units to be closely related to our discourse units. In (Ford and Thompson, 1996), the interactional unit is defined through three key criteria: syntactic completion, intonation completion, and pragmatic completion as illustrated

in (2). While prosody is not the focus here, syntactic completion relies on the syntactic structure without being restricted to the traditional notion of the "sentence". Instead, it helps to make the concept of *'projectable units of natural language'* concrete (Sacks, 1992). This relies on the 'clause' notion, where a syntactically complete utterance can, "in its discourse context, be interpreted as a complete clause, that is, with an overt or directly recoverable predicate". Interactional units also depend on identifying a prosodic final intonation contour and require the unit to be interpretable as a "complete conversational action".

(2) from (Ford and Thompson, 1996)
it was like the other day ! uh > # vera > was talking ! on the phone ! to her mom!>][4]

2.1.2 BASIC DISCOURSE UNITS

Identifying *Basic Discourse Units* is central to a series of studies by Liesbeth Degand and colleagues (Degand and Simon, 2005, 2009; Crible and Degand, 2019; Hu and Degand, 2023). More precisely, (Hu and Degand, 2023) follows the general approach of (Degand and Simon, 2009) in defining *Conversational Discourse Units (CDUs)*.[5] They emphasize the importance of treating syntax, prosody, and pragmatics as independent dimensions at the same level of granularity. Syntactic units are identified using dependency parsing (Nivre, 2010). Specifically, a syntactic unit consists of the verb along with its arguments, forming a 'dependency clause' that demonstrates maximal syntactic completeness. This approach excludes some linguistic elements (e.g., fragments without verbs, discourse markers), which then need to be promoted to syntactic units, resulting in an exhaustive segmentation of the discourse into syntactic units.[6] The pragmatic dimension is inspired by Ford and Thompson (1996) but operationalized in a way that does not depend on the interlocutor's response. A pragmatic unit is understood as a step in the implementation of a speaker's plan. However, CDUs should not be limited to the smallest syntactic, prosodic, or pragmatic units. Rather, they are the result of the integration of these three levels of analysis. Consequently, a CDU is a multidimensional unit, with several subtypes emerging depending on the specific mapping between the three dimensions.

2.1.3 MACROSYNTAX AND DUALISTIC APPROACHES

Building on the macrosyntax movement, which focuses on spoken French syntax (Blanche-Benveniste et al., 1990; Deulofeu, 2003; Sabio, 2006; Benzitoun and Sabio, 2010), Pietrandrea and Kahane (2019) proposed an operationalization for analyzing spontaneous speech utterances in terms of syntactic, prosodic, and discourse organization. The model suggests that discourse is composed of *Illocutionary Units (IUs)*, based on Austin's concept of Illocutionary Force (Austin, 1975). Each unit consists of a *nucleus* that carries the illocutionary force, along with optional elements, or *satellites*, which depend on their nucleus. These satellites are studied without directly referring to traditional syntactic categories. Additionally, the macrosyntactic framework provides specific notations to transcribe parentheticals, connectives, and other phenomena frequent in spontaneous speech. Analyzing a sequence within the macrosyntactic framework requires expertise, as it de-

4. In this example, "!" stands for syntactic completion point ; ">" prosodic completion point and ']' correspond to the pragmatic unit.
5. *Conversational Discourse Units* are conversational counterparts of *Basic Discourse Units*.
6. By 'exhaustive', we mean that the segmentation is continuous and no tokens are left behind.

mands a fine-grained understanding of the utterance. The framework uses highly specialized terms and concepts.

(3) from (Pietrandrea and Kahane, 2019)
[(je suis arrivée euh au Kenya) (je voulais travailler d'abord pour le gouvernement)]$_{iu}$
[(I arrived erm in Kenya) (I wanted to work first for the government)]]$_{iu}$

As can be seen in example (3), in this framework *Illocutionary Units* (here within brackets) can go beyond standard microsyntactic (called here Government Units here in parentheses).

More precisely, the annotation process involves several distinct tasks: (i) identifying illocutionary units; (ii) annotating the internal structure of an IU; and (iii) determining the relations between IUs. IUs can be defined as all the microsyntactic units (which relate to traditional syntactic relations, e.g., dependencies) that contribute to realizing a single assertion, injunction, interrogation, or exclamation. This is the case when they could be embedded under the scope of a saying verb that makes the illocutionary value of the entire sequence explicit. As mentioned earlier, the nucleus plays a crucial role, as it bears the illocutionary force of the entire sequence. The nucleus could be uttered alone and still retain its illocutionary function.

Several other attempts to define systems better suited to the peculiarities of spontaneous speech exist. Brazil (1995) proposes a functional grammar of speech, aiming not to describe 'sentence forms' but to describe the successful realization of communicative purposes. He uses the terms 'increment' and distinguishes between 'telling increments' and 'asking increments'. When defining the minimum requirements for a telling increment, he mentions the inclusion of nominal and verbal elements, with predication playing again a central role.

Within his dualistic approach to grammar, Haselow (2017) articulates a micro- and macro-grammar, identifying several *"fields"* (e.g., *initial*, *medial*, and *final*) to refine the modeling of spontaneous speech sequences. It should be noted that peripheral utterances have been scrutinized as key locations for analyzing spoken syntax, discourse, or conversational organization (Lewis, 2021; Herment et al., 2022).

2.1.4 FORMAL MODELS OF DISCOURSE AND DIALOGUE

Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003) is a theory that models the semantic-pragmatic interface by constructing a coherent discourse structure. In SDRT, discourse is divided into elementary and complex discourse units (EDUs and CDUs). Despite its compositional semantic foundation, the definition of *EDUs* is also top-down. Any object in the text that serves as an argument of a discourse relation can be considered a discourse unit. For the standard case, SDRT relies on the semantic notion of a *proposition*, which is closely tied to the syntactic notion of an independent clause. However, SDRT also accommodates a wide range of forms, including propositional pronouns and non-sentential units, which can be promoted to this level.

Stent (2000) made also an early proposal to extend RST (Mann and Thompson, 1987) to dialogue. They proposed a multi-level definition indicating to first segment on *'cue words'* separating *'syntactic phrases'* (involving discourse and syntax factors); then treat *'syntactically complete clause'* (syntax) and then consider *'stretch of continuous speech ended by a pause, a*

prosodic boundary or a change of speaker' (prosody, interaction).

While framed within a completely different formal apparatus, Ginzburg (2012)'s model of dialogue also allows nearly any 'form' to function as a dialogue move. To account for the broad range of forms—such as elliptical clauses, answers to questions, back-channel responses, as well as laughter or gestures, Ginzburg relies on a rich modeling of conversational context.

2.2 Natural Language and Speech Processing

When considering automatic discourse segmentation for spontaneous speech transcripts, there are two main traditions to consider: textual discourse segmentation and segmentation of speech into sentence-like units.

2.2.1 TEXTUAL DISCOURSE SEGMENTATION

As early as (Polanyi and Scha, 1983), a discourse model was proposed, even though it provided little details about its basic building blocks. The primary unit identified was the *'clause'*, a semantic object that seemed to require no further specification. Much later, Passonneau and Litman (1997) initiated a new standard of discourse processing studies by conducting an annotation campaign, including inter-annotator agreement evaluation, and developing a system for discourse segmentation of spoken data. This system combined prosodic features (primarily pause duration) and discourse connective information to approximate the human reference. However, their focus was on paragraph or topic segmentation rather than on segmenting elementary discourse units.

Overall, the most common task in textual discourse segmentation has been the segmentation of long documents based on their thematic and rhetorical organization (Hearst, 1994). This was considered an important precursor for other tasks, such as summarization or argumentation structure extraction (Marcu, 2000). Identifying elementary discourse units is sometimes seen as a necessary step for comprehensive discourse parsing (Marcu, 2000). Stede (2012) elaborates on this, emphasizing that elementary discourse segmentation is an important step in the discourse processing pipeline. However, in the context of written documents, this task is limited to the identification of sentence-internal discourse units that reside within a sentence. Stede offers a semantic perspective on elementary discourse units (Stede, 2012, p.89), defining them as: *"A span of text, usually a clause, but in general ranging from minimally a (nominalization) NP to maximally a sentence. It denotes a single event or type of event, serving as a complete, distinct unit of information that the subsequent discourse may connect to."* This definition combines both internal criteria (what constitutes a discourse unit) and external criteria (valid attachment points for subsequent discourse).

Following the general trend in NLP, symbolic rule-based approaches (Polanyi et al., 2004) have been largely replaced by statistical methods (Soricut and Marcu, 2003). As in many other subfields, Large Language Models (LLMs) have proven highly effective, achieving unprecedented performance on discourse segmentation tasks. However, discourse segmentation within deep learning approaches has been applied to only a limited number of languages, until the recent launch of the DISRPT campaigns (Zeldes et al., 2019, 2021; Braud et al., 2023). The research conducted within the framework of these campaigns has provided the community with powerful tools and frameworks to perform discourse unit (DU) segmentation and its evaluation using contemporary

methods. However, even for written genres, discourse segmentation performance deteriorates in languages other than English and when gold sentences are not provided, due to the imperfections of sentence splitters (Braud et al., 2023). Nevertheless, a recent trend involves using sequential models over contextual embeddings for discourse segmentation (Pruksachatkun et al., 2020; Muller et al., 2019). Metheniti et al. (2023) provide an improvement over (Muller et al., 2019), achieving new state-of-the-art results for discourse segmentation in various languages.[7] Several aspects of our study build upon this framework.

In addition to fine-tuning approaches, zero-shot and few-shot learning techniques are gaining significant attention across a wide range of tasks. While these methods present challenges related to explainability, control, and prediction-time costs, they constitute a new paradigm for our field. These techniques have been explored for discourse relation identification (Metheniti et al., 2024) and discourse segmentation (Nayak, 2024). The latter, which directly employed ChatGPT[8] for English discourse segmentation, found that results were unsatisfactory due to model-generated so-called hallucinations, i.e. generation of spurious elements.

2.2.2 DIALOGUE ACT IDENTIFICATION

The development of dialogue systems requires the identification and classification of user responses as communicative acts. Early work has primarily focused on dialogue act classification (Stolcke et al., 2000), often considering segmentation to be an implicit task, handled by segmenting sentence-like units (see the next section), or implicitly by whatever pre-existing given boundaries (e.g., short speaker turns typically corresponding to a single communicative act).

Gupta and Bangalore (2003) propose a pipeline that starts with the user's utterance, extracting and removing filled pauses, discourse markers, and explicit edits, followed by the identification of coordinating devices to produce what is termed the *"clausified utterance"*. Although this work primarily targeted human-computer interaction, which presents distinct challenges compared to segmenting human-human conversational speech, many steps in their pipeline align with the difficulties faced in systematically segmenting spontaneous speech transcripts.

The majority of the work in this area targets the classification of dialog acts and accordingly attempted to treat segmentation and classification as joint task (Quarteroni et al., 2011). The approaches and the corresponding papers are not very detailed with regard to the segmentation specifically. However, (Ang et al., 2005) specifically discusses segmentation evaluation metrics. Zhao and Kawahara (2018) propose a joint approach and use a BiLSTM for encoding and a sequence classifier for decoding the segmentation part of the model, evaluated on the Switchboard data set (SWDA, Jurafsky et al., 1997). Some studies attempted to work directly from the speech signal such as (Dang et al., 2020) that integrates the dialogue act segmentation with the ASR through the use of a *'dialogue act boundary'* token.

Finally, dialogue act tagging, due to its specific application context, gives rise to varying interpretations of the base units. For instance, some authors argue that dialogue acts are multi-functional, and thus multiple segmentations may be considered depending on the aspect of the dialogue being

7. Code available at `https://github.com/phimit/jiant/`
8. `https://openai.com/research/gpt-4`

analyzed at the time of segmentation (Bunt, 2011). See example (4) for a case of discontinuous functionality with the specific *time-management* function interrupting the main communicative function.

(4) from (Bunt, 2011)
[The first train to the airport on Sunday is at # [let me see] # 5.32]

2.2.3 SENTENCE SEGMENTATION IN SPEECH

The emergence of large conversational corpora (Godfrey et al., 1992) introduced several challenges for the speech processing community, with one of the most significant being the unruly nature of conversational speech flow. To perform any kind of processing, it was necessary to segment this flow into convenient units. The decision was made to segment it into sentence-like units, with various perspectives regarding their relation to written language. The work done in this area, with notable success, can be summarized as addressing how to use language models (initially simple n-grams) for predicting sentence-level breaks in the token flow, how to incorporate acoustic-prosodic information, and how to combine these two sources of information to improve segmentation results.

In terms of language modeling, early models based on Hidden Markov Models (HMMs) (Stolcke and Shriberg, 1996) already produced promising results. One difficulty they faced was the strong impact of disfluencies in the natural speech flow, leading to the development of specific models designed to identify and eliminate them (Stolcke et al., 1998). Over time, these models were improved by adopting Conditional Random Fields (CRFs) (Liu et al., 2005). These models aimed for efficiency but imposed a strong bias from the written realm on any subsequent data analysis.

Several studies have shown the benefit of considering acoustic modality characteristics in speech segmentation. Building on the pioneering studies of Pierrehumbert and Hirschberg (1990) and Hirschberg and Grosz (1992) about the relationship between prosody and discourse structure, Shriberg et al. (2000) systematized the use of acoustic-prosodic cues for segmenting speech into sentence-level and topic-level units. Furthermore, they discovered that the importance of different cues varied depending on the corpus considered. For instance, pause and pitch features were highly informative for segmenting news speech, while pause duration and language modeling dominated in natural conversation. See also (Portes and Bertrand, 2011) for an investigation of the issue in French.

Approaches based on low-level acoustic features, such as pauses, have since been the primary method for segmenting speech into utterances. However, the performance of these simple systems plateaued over time, and with the rise of more advanced machine learning techniques, methods trained on much larger written corpora (allowing for language modeling at a new level) have been explored as effective alternatives. One of the most direct approaches to segmenting speech transcripts into communicative units is to punctuate them as if they were written. The general strategy involves training a punctuation model on a large corpus of written data and then applying this model to speech transcripts lacking punctuation, as done by Favre et al. (2008). This method provides an efficient approximation of segmentation into discourse units. It works particularly well for fluent, canonical, and monological sequences, where the method is nearly perfect. However, such models struggle with disfluencies and highly spontaneous sequences absent from the training corpus. More-

over, from a linguistic perspective, it is questionable to impose a notation derived from the written realm onto these spontaneous speech transcripts.

Despite these limitations, punctuation-based methods remain efficient, with newer models employing multilingual deep language models, e.g., XLM-ROBERTA, which has proven to be the best-performing model (Guhr et al., 2021).[9] While these models provide impressive results for a relatively simple approach, they note that "punctuation patterns are domain-specific, and robust punctuation prediction requires training on diverse datasets."

2.3 Take-away

From this literature, we retain the following main take-aways. Overall, identification of discourse units in spontaneous speech is a complex problem in which syntactic, prosodic and pragmatic / interactional dimensions are deeply intertwined. Many proposals (Ford and Thompson, 1996; Degand and Simon, 2009; Petukhova et al., 2011; Hu and Degand, 2023) are going as far as treating this issue by considering units at all those different levels and discourse units as products of the relationship between these different sub-level units.

A second idea well represented in these works is the need to adapt the grammar itself for the internal structure of the discourse units in spontaneous speech. The main systematic issue is the presence of disfluencies. Overall, spoken specific structure of discourse units are built around a more linear organization around a central element and a range of optional initial and final elements. These organizational patterns may offer a way to bridge the 'local' and 'global' levels of discourse, with the peripheries naturally serving as predisposed locations for anchoring elements at the global level.

The ultimate criterion shared by many seems to be the discourse function / communicative action. This corresponds to the idea that discourse segmentation and discourse relation analysis should not be performed in a strictly sequential way (Hoek et al., 2018). Roulet et al. (2001) go as far as considering that the minimal discourse unit definition must be a top-down one as there is no criteria to define a syntactic maximal unit. Overall however, even though its individuation is recognized to be difficult, the different accounts rely on the core notions of predication, semantic proposition and speech act.

In terms of automatic segmentation, it seems that the convergence of NLP (nowadays represented by LLMs, pretrained on huge amounts of textual data, written or conversation) and speech processing traditions, that also used language modeling coupled with acoustic extraction is not completely achieved. There is currently a convergence in terms of methods but the segmentation of elementary discourse units in spontaneous speech has not yet greatly benefited from this convergence. One explanation is that overall NLP and Speech Processing are moving toward more end-to-end approaches that avoid processing pipelines in which discourse segmentation would play a role.

9. See also `https://github.com/benob/recasepunc` for alternative implementations.

3. Operationalization: Definition and Examples

Based on this review, we adopt a discourse approach that we aim to keep simple enough for semi-naive coders after receiving some training. While acknowledging the multidimensional nature of the phenomena discussed, our approach focuses on a single level of discourse/semantic/pragmatic units. Even though multiple overlapping layers are analytically justified, we believe that conversation participants and semi-naive coders should identify discourse units as conversational actions without explicitly attending to these complex overlapping layers.

Our operationalization of elementary discourse units employs both top-down and bottom-up characterizations. This mirrors the interface position of EDUs as both the maximal unit within the realm of grammar and language and the minimal unit of discourse structure, including in texts, stories, narratives, argumentation, or task-oriented dialogues.

From the top-down perspective, any communicative object with, or promoted to have, propositional-level value (including non-assertive values such as questions or requests) that can be related to another discourse object through a discourse relation is considered a discourse unit.[10] Our approach will group semantic, pragmatic, and discourse aspects under the common-sense ideas of ”what it means” or ”what it does” in context.

From the bottom-up perspective, we also rely on syntax, though without limiting ourselves to independent clauses. This flexibility is enabled by the top-down criteria and informed by the variety of phenomena and structures identified in the literature on spoken syntax and spontaneous speech. We adopt a generous approach to what can be included in a discourse unit, drawing on literature on ”satellites,” ”fields,” and ”peripheries” (cf section 2.1.3). This approach consists in allowing for these elements to be included in, potentially large, elementary discourse units as long as they do not clearly convey a full propositional meaning and are not discourse-articulated (e.g. with a discourse marker) to the rest.

3.1 Definition

The discourse units we aim to annotate are closely related to the ones of (Polanyi et al., 2004), where they communicate information about no more than one event, event-type, or state of affairs. These units correspond semantically to independent syntactic clauses. However, the interactional and dialogical nature of discourse requires us to expand this view to include any object that, in context, plays a similar semantic role. Our semantic perspective on discourse units leads us to combine criteria to reach the definition proposed further in the definition below.

- **Semantic criteria:** Following Vendler (1957)’s classification of eventualities, we consider a discourse unit as a segment representing a specific event, state, or proposition ;

- **Discourse criteria:** We are using discourse markers, whether these markers are connectives, bundles of them or adverbials in a discourse position.

10. We do not elaborate on the notion of discourse relations here as it would take us too far from the segmentation focus. It is enough to mention that discourse units can be related to or attached to other units from the discourse context.

- **Pragmatic criteria:** We recognize specific speech acts, such as asking / answering a question or providing conversational feedback, as discourse units.

Definition *In a speech transcript, an elementary discourse unit is a span of contiguous tokens whose combined meaning represents a semantic proposition (describing an event, fact, or state of affairs) or an individual speech act, which can be identified as a basic communicative function, such as asking a question, providing an answer, or offering conversational feedback. All preparatory disfluent material should be included within the elementary discourse 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.*

To illustrate, a discourse unit is a segment that describes an eventuality, as in example (5), or carries a clear communicative function, as seen in speech acts such as in (6). In the former case, identifying a predicate (e.g., through the main verb) is a strong clue, as discussed in the related work section. The discourse unit generally includes the arguments of the verb unless strong discourse markers signal a distinction between the arguments and the rest of the clause, as in example (7).

(5) 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 who hit a stone bench]$_{du}$

(6) 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]$_{du}$

(7) Discourse Markers inducing segmentation

a. [on a appelé euh des les parents d'amis]$_{du}$
[we called um some friend's parents]$_{du}$
[mais pas d'amis de notre âge d'amis de mes 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]$_{du}$

Discourse adverbials and conjunctions serve as strong cues for discourse boundaries and are included in the discourse unit they introduce, as illustrated in example (7). Final spoken particles like

en fait, quoi (in fact, what) also provide useful cues for discourse segmentation and are included in the preceding unit.

Strictly speaking, attitudinal markers such as *tu sais* (you know) or *je crois* (I think) can be seen as discourse units. However, we instructed annotators to include them within the units they qualify. First of all, the discourse attachment of these elements is rather trivial : they qualify the discourse unit within which they occur. This decision also stems from the fact that these markers are often produced as parentheticals, and treating them as standalone discourse units would result in many nested discourse units, which we aim to avoid. Furthermore, in terms of methodology, these attitudinal markers are easier to treat specifically at a later stage.

3.2 Data

The experiments presented in this paper use an existing discourse-segmented corpus, which was segmented according to the approach outlined above. The raw data comes from the *Corpus of Interactional Data (CID)* (Blache et al., 2009, 2017), which consists of 8 dyadic conversations, each lasting approximately 1 hour. The CID contains highly spontaneous data, featuring colloquial sequences, as in example (8), and strong disfluencies, as in example (9), making discourse segmentation much more challenging than for written genres, even for human annotators.[11] The full dataset consists of approximately 125,000 tokens and 15,463 discourse units, with 12.4% of the tokens marking EDU boundaries.

Annotations were performed using Praat (Boersma, 2002) to enable precise word-level alignment of the audio signal when making segmentation decisions. More specifically, during EDU segmentation, annotators had access to the time-aligned transcripts of both participants in the conversation.

The discourse annotations presented and used in (Prévot et al., 2021) were carried out during the OTIM project (Blache et al., 2009, 2017) and are available on the Ortolang platform, along with the guidelines used for annotation.[12]

(8) A:[comme ça # ah ouais non c'était]$_{du}$
[like that # oh yeah no it was]$_{du}$
B:[ah ouais profitez profitez de vos soirées]$_{du}$
[oh yeah enjoy enjoy your evenings]$_{du}$
A:[ouais c'est pour ça]$_{du}$
[yeah like that]$_{du}$

(9) [ou des euh non pas des f- pas des frustrations]$_{du}$ [des # espèces de euh # mhm # ouais des des vues différentes sur le boulot quoi]$_{du}$
[or some uh no not some f- not some frustrations]$_{du}$ [some kind of uh # mh # yeah some some different views about work what]$_{du}$

11. We use the symbol '#' to mark pauses in the examples.
12. https://www.ortolang.fr/market/item/ortolang-000918

3.3 Manual Segmentation Evaluation

Elementary Discourse Units were obtained through at least two manual annotations, conducted by 4 naive coders and 2 experts. The mean Cohen's κ score across speakers for naive coders is 0.85 (min: 0.83, max: 0.87). For evaluating multiple coders' agreement we used a standard multi-κ measure discussed in (Artstein and Poesio, 2008). More specifically, we additionally report, in the appendix, the multi-κ score when available (Table 9), the agreement between individual naive coders (Table 8), and the agreement between naive coders and experts (Table 10).[13]

While we acknowledge the advantages of using adapted metrics such as *WindowDiff* (WD) (Pevzner and Hearst, 2002), *Boundary Edit Distance* (Fournier and Inkpen, 2012), or γ (Mathet et al., 2015) — and have experimented with them in previous work (Peshkov and Prévot, 2014; Prévot et al., 2016) — the present task is a simple binary decision over a skewed, but not highly skewed, distribution. We therefore do not evaluate the task as an interval labeling problem but rather as a token labeling task, following the current trend adopted, for instance, in the DISRPT shared tasks (Zeldes et al., 2019; Braud et al., 2023).

4. Automatic Discourse Segmentation Experiments

In this section, we present a series of experiments aimed at automatically segmenting the CID corpus into elementary discourse units (EDUs) as defined earlier. The primary focus of our experiments is on fine-tuning approaches, using large language models (LLM), specifically XLM-ROBERTA-LARGE (Conneau et al., 2019). This model is on the lower end of current models with respect to size (and thus expressivity), but was chosen as it presents a good compromise between performance and ease of use without large computing resources. It is also one of the rare multilingual models in this capacity range. We compare two distinct fine-tuning strategies: a straightforward fine-tuning approach and one based on weakly supervised annotation, leveraging less annotated data. Our models are sequence-to-sequence models, fine-tuned following the methodology proposed by (Metheniti et al., 2023) within the JIANT framework (Pruksachatkun et al., 2020).[14] Additionally, we explore zero-shot and k-shot in-context learning techniques with larger generative models.

The task is defined within the DISRPT framework. The 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 1). Evaluation is performed using F-score, precision, and recall, calculated specifically for the discourse boundaries (with true negatives excluded from the score). For written data, on French, scores are around 90 of F-score (Braud et al., 2023). While there have been few attempts at this task for conversational speech, Gravellier et al. (2021) achieved an F-score of 73.7.[15]

4.1 Weak Supervision

Modern NLP techniques require large amounts of annotated data. The general idea behind weak supervision is to avoid the costly process of manually annotating large datasets by using

13. The agreement between experts is not meaningful, as expert double-coding was performed only on a small subset of the data, which was also used for training the naive coders. As a consequence, expert agreement is nearly perfect, as it underwent several rounds of adjudication. However, we obtained an expert agreement of about 0.9 for the same task on a different corpus (Prévot et al., 2025).
14. See `https://github.com/phimit/jiant-discut`
15. At the time, the base LLM used was BERT. Using XLM-ROBERTA would likely result in a higher score.

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 1: Illustration of the data format for the discourse segmentation task.

semi-automated methods to generate annotated data. While the accuracy of these annotations is known to be lower than human-generated annotations, the large volume of data is expected to compensate for the noise. More precisely, we adopt here the "data programming" approach of Ratner et al. (2017), using the Skweak framework developed by Lison et al. (2021).[16] This framework provides an API for defining multiple overlapping heuristic classification rules and an aggregation model to combine them. The set of rules, called labeling functions (LFs) in that framework, can be developed and tested with only a small amount of annotated development data. The system builds a profile for each LF, and a model is trained by combining all the LFs, which can be weighted by their estimated reliability. The reliability is estimated without supervision, relying on agreement between rules, and the hypothesis that they are all at least slightly better than pure chance. This "label model" is then used to label a training set, and finally, a supervised model is trained on the dataset automatically annotated by the label model. A similar approach, based on the SNORKEL implementation of (Ratner et al., 2017), has been used in discourse analysis to enrich a discourse parser (Badene et al., 2019), and is also the foundation for the work in (Gravellier et al., 2021).

In our experiments, we combine three types of rules: (i) base rules, which assign a label (positive or negative) to almost all tokens (acting as a default rule), which can come for instance from a model trained on written text; (ii) recall-oriented rules, which aim to catch more specific discourse boundaries, and (iii) precision-oriented rules, which specify where discourse boundaries should not occur (e.g., by narrowing the application scope of the default rules). For instance, in Table 2, the rule `lf_du_non_end_tok` labels tokens that are unlikely to mark the end of a discourse unit, while the rule `lf_long_pause` labels tokens following a pause of at least 800 milliseconds.[17]

We created four sets of rules for weak supervision, based on two key considerations: (1) variations of the base rule, either relying on the DISCUT model (Metheniti et al., 2023) or on a sim-

16. https://github.com/NorskRegnesentral/skweak

17. A few examples of labeling functions are provided in Appendix C, while the complete set is available in the repository at https://github.com/phimit/WeakLing/tree/main/weaksupervision.

#	name	label	conflicts	precision	recall	F-score
8	lf_du_non_end_tok	NDU	0.017	0.996	0.359	0.528
9	lf_du_non_end_pos	NDU	0.027	0.990	0.427	0.596
14	lf_du_very_long_pause	DU	0.061	0.948	0.246	0.390
17	lf_du_long_pause	DU	0.150	0.868	0.378	0.526
19	lf_du_dm_ini_forward	DU	0.347	0.683	0.113	0.194
20	lf_du_discut	DU	0.410	0.607	0.876	0.717

Table 2: Some discourse segmentation labelling function profiles. Conflicts : % cases conflicting with at least another labelling function. DU : Discourse Unit Start = True ; NDU : Discourse Unit Start = False. As an example rule #14 says that a very long pause implies the start of a new segment, and rule #9 says that at a token from a given list of part-of-speech tag, there is no boundary at that token. Precision, recall and F-score are estimated from development data to help rule design, but are not taken into account by the weak supervision process.

ple punctuation predictor[18]; (2) whether to use syntactic chunk information (mostly as precision-oriented rules to prevent splitting chunks into distinct discourse units) or not. These variations led to the following weak supervision models being tested:

- `mono-discut`: Simple discourse segmentation fine-tuning (FT) with DISCUT as the base rule for weak supervision.
- `multi-discut`: Joint segmentation and chunking with DISCUT as the base rule for discourse noisy labels (and chunks also having noisy labels).
- `mono-punct`: Simple discourse segmentation FT with a PUNCTUATION PREDICTOR as the base rule for weak supervision.
- `multi-punct`: Joint segmentation and chunking with a PUNCTUATION PREDICTOR as the base rule for discourse noisy labels (and chunks also having noisy labels).

Adding chunk information did not seem to improve discourse segmentation performance, however. As a result, we focused on sets of rules that did not incorporate chunk information when predicting noisy labels. Additionally, we conducted experiments both with and without DISCUT, as this represents two realistic scenarios for different languages, depending on whether a high-quality segmentation model exists for the language in other genres.

We compared our models with the set of models that are either baselines, toplines or interesting competitors. All the models are fine-tuned models based on XLM-ROBERTA-LARGE:

- `gold` : model fine-tuned on the training set with gold labels for the whole training set **(topline)**;

18. We also experimented with pauses as the base rule, but this resulted in acceptable, yet significantly lower, performance compared to the other approaches.

model	precision	recall	F-score	nb supervised tokens	nb supervised DU
fra.gold	0.865	0.830	0.847	115850	15630
fra.gold10	0.831	0.839	0.835	11585	**1563**
fra.mono-discut	0.698	0.850	**0.767**	11581	1498
fra.multi-discut	0.671	**0.876**	0.760	11581	1498
fra.mono-punct	0.776	0.684	0.727	11581	1498
fra.multi-punct	**0.797**	0.662	0.723	11581	1498
fra.discut-w	0.604	0.872	0.714	0	0
fra.pause	0.896	0.423	0.575	0	0
fra.punct	0.776	0.443	0.564	0	0

Table 3: Results of experiments on the French conversation corpus with the main approaches and a few toplines and baselines. GOLD is a fully supervised model as a topline comparison, GOLD10 the same with only 10% of the annotated data. DISCUT_W is the DISCUT model trained on written text and taken as it is for conversation segmentation. PAUSE and PUNCT are baselines relying only on pauses or automatically predicted punctuation respectively. In bold are the best weak supervision scores. The number of supervised tokens/DU needed for each approach means annotated instances for training a fully supervised model, or in the case of weak supervision, the number of annotated instances used for development/validation of the heuristic rules. This was predetermined to be roughly the same as 10% of the gold annotations, but ideally a weak supervision approach might need less development instances for similar results.

- `gold_10` : same model fine-tuned on only 10% of the gold-annotated training set (which corresponds roughly to the amount of gold data used to produced the weakly supervised development set);
- `discut_w` : discut model directly used to predict on our data set (no fine-tuning) **(baseline)**. Pauses longer than 500 ms were sent to the model as commas;
- `pause` : Rule-based model segmenting the data set on pause over $500ms$ **(baseline)**;
- `punct` : Rule-based model using strong punctuation (period, question mark) as predicted by the punctuation prediction model **(baseline)**.

The main result of this set of experiments, presented in Table 3, is that, within this framework and for this language, fine-tuning on a corresponding amount of gold labels is more efficient than adopting a weakly supervised approach where gold annotations are used to develop and evaluate the heuristics rule. This amount has to be predetermined before developing the rules, and it is possible that less annotations are needed, but it is hard to estimate, and the amount of gold annotations used in our experiment is actually not that big (about 11k tokens and 1500 discourse units). This contrasts with previous results from (Prevot et al., 2023), where simple fine-tuning required up to 7000 discourse units (DUs) to outperform weak supervision. The improvement observed here can be attributed to the use of ROBERTA (compared to BERT), the DISCUT framework (compared to TONY), or both combined. While it is possible that better labeling function engineering could yield

improved results, we have experimented extensively with the available cues and accumulated experience on this issue (Prevot et al., 2023). While not completely ruling out this possibility, in terms of efficiency, dedicating time to developing the labeling functions does not seem more efficient than performing annotation on a sample. Moreover, we observe that using DISCUT (a discourse segmenter trained on written data) yields better results than a "simple" punctuation predictor. Finally, we do not observe a clear difference between the results obtained in the single-task versus the multi-task approach. One open question is that heuristic rules might be more robust to change in the distribution of data they are applied to, but this goes beyond the present study.

4.2 How much human-labeled data do we actually need?

The previous experiment, as well as the work conducted in (Prévot and Wang, 2024) for another language, led us to explore more systematically and radically how much data is actually needed to create an effective segmentation model. We used again XLM-ROBERTA 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. We performed 8-fold cross-validation based on speaker IDs, creating different test, development, and training splits to maximize the distance between training and testing data given our corpus.

nb train DUs	nb train tokens	precision	recall	F-score
≈ 15	100	0.73 ± 0.29	0.11 ± 0.06	0.19 ± 0.09
≈ 30	200	0.81 ± 0.11	0.21 ± 0.08	0.33 ± 0.10
≈ 70	500	0.77 ± 0.03	0.63 ± 0.04	0.69 ± 0.03
≈ 140	1000	0.78 ± 0.04	0.73 ± 0.02	0.75 ± 0.03
≈ 700	5000	0.80 ± 0.02	0.80 ± 0.05	0.80 ± 0.02
≈ 1400	10000	0.82 ± 0.02	0.81 ± 0.05	0.82 ± 0.02
≈ 4200	30000	0.82 ± 0.02	0.84 ± 0.04	0.83 ± 0.02

Table 4: Global simple Fine-Tuning Results with different amounts of training supervision, averaged with leave-one-speaker-out cross validation (resulting in 8 folds).

Figure 4 summarizes the results of our experiments. Fine-tuning ROBERTA for elementary discourse unit segmentation proves to be highly efficient, even with a small amount of training data. The general pattern observed is that the model starts with a very low recall score but improves quickly as the amount of training data increases. Answering the question in the title of this section depends of the precise goal of the discourse segmentation. If one simply wants more accuracy than the usual proxies, then we can seen that even a modest training build within a few days can be enough. However, if one is looking for the best discourse segmenter possible with the goal to approach human level, more data is desirable. The F-score tends to plateau but still goes up when training size grows exponentially.[19]

19. See however (Prévot et al., 2025) for related experiments on a different corpus but with more training data and a more pronounced "plateau" effect.

4.3 Comparison with few-shot generative approaches

Given the dynamics of the fine-tuning experiments presented above, we decided to explore the potential of a zero- or few-shot approach using a generative model for this task. We initially attempted a pure zero-shot approach using the prompt in Example (10), and then adopted a few-shot approach with the prompts provided in Appendix D (see example (11) for a short illustration). The few-shot prompt consists of a set of examples derived from our corpus, taken from the training portion. For each example we provide the input data without the segmentation and the expected output using the "|" pipe symbol as a boundary label.
Note that we pre-tested both an English prompt and the same prompt in French, and the former seemed to work better.

(10)
```
"Segment the following dialog in elementary discourse units,
where the character # indicates a short pause:  <text>
Only print the original text, indicating segment boundaries
with a | character, and do not add anything, do not remove
anything.  Do not present the result with introductory
text."
```

We used a version of the freely available Llama3-8B base model (Dubey et al., 2024), quantized to 4 bits, served via the Ollama wrapper and API.[20] The model was set with a temperature of 0.3 and top_k of 20 with the idea of generating conservative outputs. To test for stability, we ran each experiment 5 times, randomly selecting examples for the prompt from our base of 10 examples. This kind of model is light enough to be run on simple hardware.

(11)
```
Example 1:
input:  ah mais c' est bon je sais plus quoi dire là c' est
bon hein i- t- # enfin je sais pas
output:  | ah mais c' est bon | je sais plus quoi dire là |
c' est bon hein | i- t- # enfin je sais pas
```

Since the generative model impacts the tokenization (adding tokens in particular), we post-processed the output with simple rules and excluded pauses from the evaluation, since it is the most added token, but never mark the beginning of a segment. This adjustment helped mitigate biases introduced by the deviant tokenization. Note that it is a supplementary step that can increase development time much more than the term "few-shot" might suggest, and indicates how brittle a generative approach can be for token-level predictions instead of simple classification, at least with moderately sized models.

As can be seen in table 5, the 5-shot experiment amounts to about 20 to 30 discourse units that correspond roughly to our smallest training size in our fine-tuning experiment.

The first conclusions we can draw from this experiment is that the generative approach works surprisingly well given how little supervision it receives, compared to fine-tuning with a few examples (lines 1-2 of table 4), but it is quite far in performance from more reasonably supervised approaches (supervised or weakly supervised), as the scores plateau quickly after a few examples.

20. https://ollama.com/

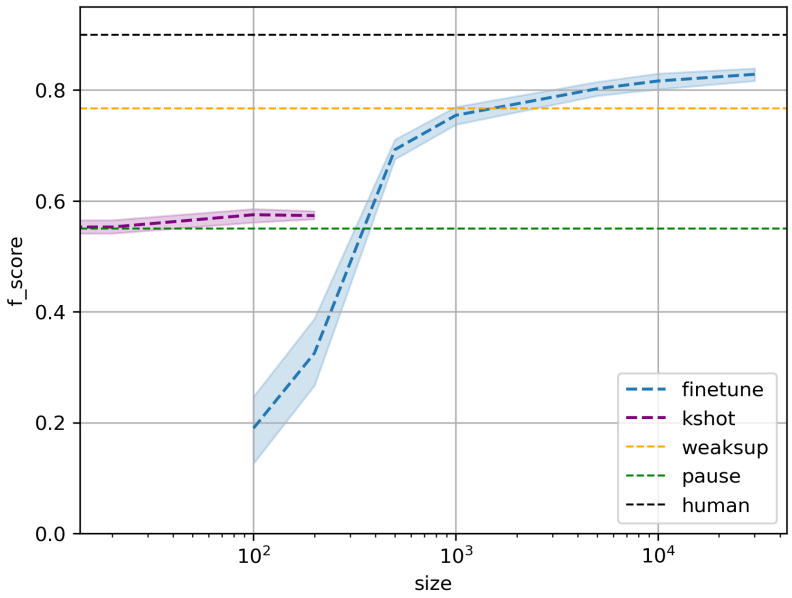


Figure 1: F-score comparing fine-tuning and few-shots approaches. x-axis: number of tokens, log scale. Green lines correspond to 200ms pause baseline and black line to human annotation topline (average across the languages).

This approach is prone to some instability by nature and could be investigated more (including with larger models), but it would lose the point of decreasing engineering efforts.

4.4 Discussion

The key finding of our experiments on French conversations is that simple fine-tuning on gold labels performs well for elementary discourse unit segmentation. The results from the zero/few-shot pilot study, however, yielded contrasting outcomes. Few-shot performance was achieved with 20-30 DU examples, corresponding to scores obtained with 200-500 training tokens using the fine-tuning approach (roughly 30-70 DUs). On one hand, this suggests that for extremely small datasets (or in cases where no gold data is available), prompting could be an option, for instance for scenarios in which pause duration (our baseline) is not available. On the other hand, the scores and error analysis indicate that fine-tuning with a few hundred annotated DUs is generally the better strategy,

Nb of examples	precision	recall	fscore
0	0.47 ± 0.01	0.58 ± 0.01	0.52 ± 0.01
1	0.51 ± 0.03	0.60 ± 0.01	0.55 ± 0.02
5	0.53 ± 0.02	0.63 ± 0.03	0.58 ± 0.02
10	0.52 ± 0.01	0.63 ± 0.01	0.57 ± 0.01

Table 5: Results for the zero- / in-context k-shot approach.

	manual	predicted
mean	7.36	7.45
std	5.46	5.76
median	6	6
mode	1	1

Table 6: Comparison of DU length distributions, manual vs. automated.

as it produces higher scores. In contrast, prompting results are unpredictable, making it difficult to plan improvements in a rational, consistent manner.

In a broader context, as discussed in Section 2, this paper provides a global view of the available strategies for elementary discourse unit segmentation on spontaneous speech transcripts. These strategies include: fine-tuning on human gold annotations (with varying amounts of training data), weak supervision (similar to fine-tuning but using automatic annotations, such as those obtained through data programming), and zero/few-shot approaches. We observe that direct fine-tuning on gold data is the most efficient solution. The amount of data required to reach an 0.8 F-score corresponds to only a couple of days of manual annotation effort, whereas setting up a weakly-supervised approach would require at least the same amount of time with similar or lower results. While the zero/few-shot approach is faster to set up, it yields lower and, more importantly, unpredictable results, which makes it difficult to plan for improvements.

As empirical linguists, our primary motivation for building a discourse segmentation tool is to enable access to large amounts of discourse-segmented data, with a comparable quality level, and to draw similar observations as if the dataset had been manually segmented, an option not available for large datasets. The standard metrics presented in Table 6 reveal that the distributional shapes of discourse unit lengths are nearly indistinguishable between the automatic and manual segmentation. This is further established by plotting the distribution lengths between the manual and automatic version of the test set, as in figure 2.

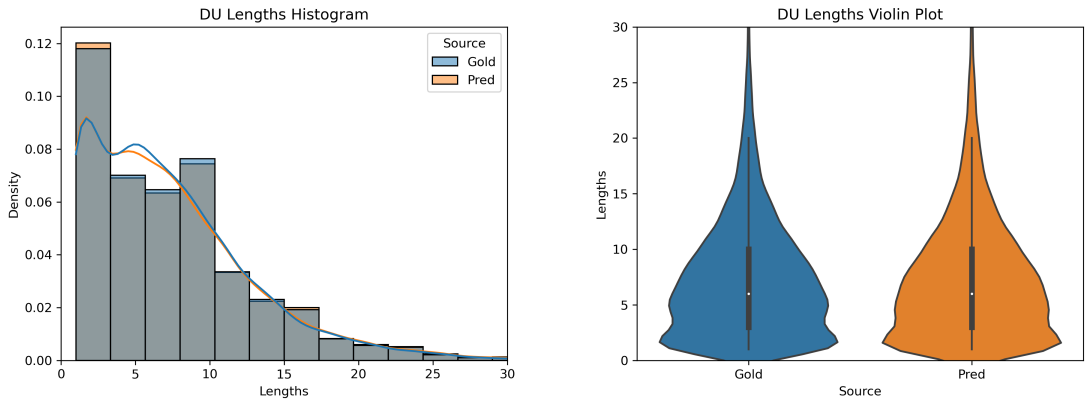
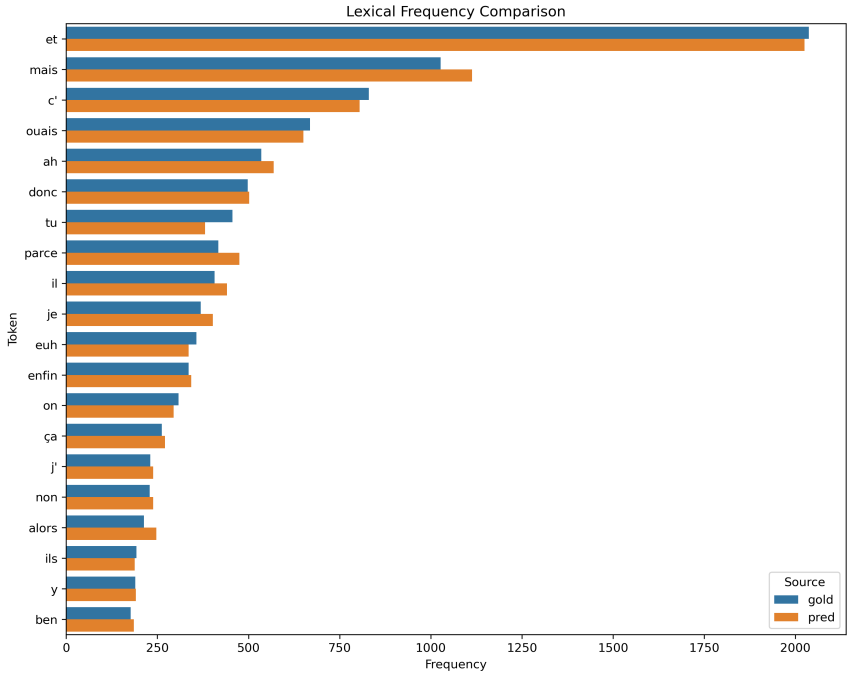
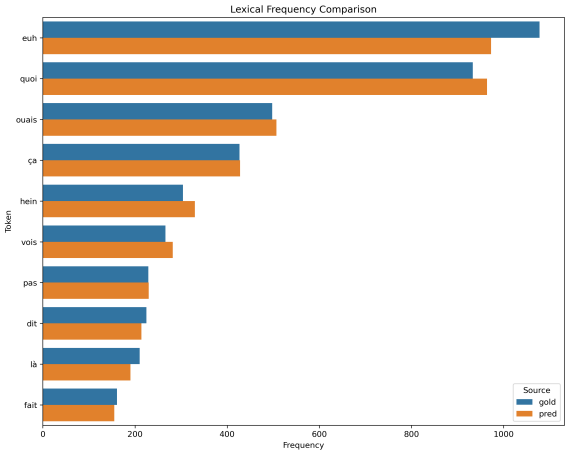


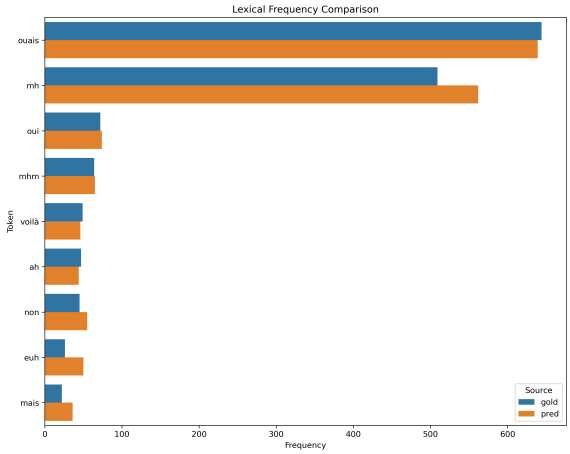
Figure 2: Distribution of DU lengths in manually and automatically segmented datasets / Dispersion of the DU lengths in manually and automatically segmented datasets in a violin plot



(a) Initial



(b) Final



(c) Isolated

Figure 3: Frequency distributions from left to right for DU-Initial, DU-Final (both for DU having at least 3 tokens) and DU-Isolated (DU made of only 1 token).

We also plotted the distribution of most frequent lexical items in crucial positions, namely DU-initial, DU-final and DU-isolated as presented in figure 3. These lexical distributions around DU boundaries, obtained automatically or manually, are remarkably similar. Some divergences occur at lower frequencies, such as the swear word *putain*, which the model does not treat as a single-token DU (despite being relatively frequent as a single DU in the manual annotation).

Type of error	
Disfluencies	34 %
Discourse Markers	27 %
Non-Sentential Utterances	8 %
Other Spoken Genre structures	8 %

Table 7: Type of errors observed in the systematic error analysis

5. Error Analysis

The scores achieved by our systems are promising but still not at the level obtained on written genres. To better understand the shortcomings of our model and identify what needs to be improved to achieve human-level segmentation performance, we conducted an in-depth error analysis. We reviewed a sub-corpus accounting for about 5000 tokens[21] from 10 different speakers in the corpus and systematically labeled the 230 segmentation discrepancies occuring in this sample between the human and automatic segmentations. We categorized errors at multiple levels, including:

- The general nature of the discrepancy between gold labels and predicted labels: *false positives*, *false negatives*, and *gold errors*.
- A general category of the cause of the errors: *disfluencies, discourse markers, pauses, non-sentential units, other spoken structures, dialogical sequences, reported speech, relative clauses*.
- The potential role of pause presence/absence in errors: *pause* (for false positives) and *no pause* (for false negatives).
- A finer-grained description of the causes of the errors: *abandoned units*, *discourse marker clusters*, *discourse marker position inversion (final vs. initial)*, *parentheticals*, *response insertions*, etc.

This analysis yielded 230 discrepancies at the token level between gold vs. predicted labels. Their systematic characterization is presented in Table 7. Interestingly, 23% of the errors were found in the gold standard itself.[22] Additionally, 48% of the discrepancies were false negatives (missed boundaries), and 29% were false positives (spurious boundaries).

As seen in Table 7, the main cause of errors were predominantly *disfluencies* and *discourse markers*. Other significant sources of error included spoken genre structures (16%), particularly *non-sentential utterances* (8%). The remaining errors were attributed to factors such as long pauses, dialogical sequences, relative clauses, or reported speech. These causes of errors are exemplified and further analyzed in the following subsections. We further examined the impact of pauses on

21. The size was arbitrarily chosen as a hopefully representative sample.
22. The manual segmentation was performed by semi-naive coders. The double-checking of these discrepancies was done by a discourse expert, co-author of this paper. This suggests that the overall system scores could be significantly higher than reported if the test set had been segmented and double-checked by experts, rather than relying on a semi-naive annotation campaign. See also (Nahum et al., 2025) for a discussion on the potential underestimation of model performance due to human errors in evaluation data.

these errors. Moreover, we found that 63% of the missed boundaries were due to the absence of a pause that the model could use, while 35% of the spurious boundaries were influenced by a pause, which likely contributed to the segmentation decision.

5.1 Disfluencies

Simple disfluencies mostly did not cause issues for the segmentation models. Errors appear when the disfluency is combined with a pause at a potential discourse break like (12) in which the sequence of a discourse marker + filler + pause triggered a spurious discourse boundary.

(12) **Disfluencies + pause misused**
[je venais de temps en temps]$_{du}$ # [et euh #] !! [et puis un jour c' était je sais pas l' été je crois]$_{du}$[23]
[i was coming from time to time]$_{du}$ # [and um #] !! [and then one day it was I don't know summer I think]$_{du}$

Complex paradigmatic piles (Gerdes and Kahane, 2009) generated some errors like in (13), maybe because the material accumulated before the decision point started to accumulate to form an acceptable looking discourse unit while the material after definitely constituted a valid discourse unit in itself.

(13) **Disfluencies + paradigmatic pile**

[parce qu' elle était pas très] !! [elle était ancienne]$_{du}$ [donc ça pouvait pas le faire]$_{du}$
[because it was not very] !! [it was old]$_{du}$ [so it could not work]$_{du}$

Disfluencies involving discourse markers were also an issue, like in examples (14): on top of the repetition of the initial discourse marker there is a pause inserted to further mislead the model.

(14) **Disfluencies + discourse markers**

a. [si euh #] !! [si tu veux je vais pas tourner de l' oeil]$_{du}$
[if uh #] !! [if you want I won't pass out]$_{du}$
b. **Disfluencies, self-edit with discourse marker**
[qui vient te euh d- #] !! [enfin bon te recadrer quoi je veux dire]$_{du}$
[who come to you uh d- #] !! [well to put you in your place I want to say]$_{du}$

Disfluencies sometimes come with truncations that results in infrequent tokens for the model and that seems to have been an issue like in (15) in which in addition to the truncation at the boundary, there is no pause before to help the segmentation decision.

(15) **Disfluency / Truncation**
[mais # mais # si si il y était justement]$_{du}$ [$$] [i- il avait obligé à le à le laisser]$_{du}$

23. In the error analysis examples, a "]!![" indicates a spurious boundary, while "[$$]" indicates a missed boundary.

[but # but # yes yes he was in finally]$_{du}$ [$$] [i- he forced to let him]$_{du}$

To conclude about disfluencies, our initial discourse segmentation model used the notion of abandoned units (see also Pallaud et al., 2013): false starts that are completely abandoned, and impossible to relate to the finally produced utterances. These abandoned units can create issues either because they are coded as separate units that the model would either include within the gold discourse unit at the beginning, as in example (17), or the end, as in example (16) of the discourse, leading to missed boundaries.

(16) **Disfluencies, Abandoned units on the right**
[genre des gens qui étaient au même niveau que moi quoi]$_{du}$ [$$] [qui étaient euh #] !! [ou qui euh qui qui l'étaient]$_{du}$ [$$] [mais qui euh de par leur âge # tu vois]$_{du}$
[like some people that were at the same level with me]$_{du}$ [$$] [who were uh #] !! [or who uh who who were]$_{du}$ [$$] [but who uh from their age # you see]$_{du}$

(17) **Disfluencies, Abandoned units on the left**
[enfin si tu veux je normalement je de- enfin]$_{du}$ # [$$] [si tout va bien je vais essayer de le faire]$_{du}$
[well if you want I supposedly I de- well]$_{du}$ # [$$] [if everything goes well I will try to do it]$_{du}$

5.2 Discourse Markers

Discourse markers, being key cues for discourse boundaries, are involved in many errors. The main reason for this is that some of the most frequent discourse markers (such as *bon* / well) can appear both at the beginning and the end of a discourse unit, depending on their function. For instance, in example (18), the discourse marker is considered to be final, while the manual annotation treats it as initial. Conversely, in example (19), the composite marker *mais bon* (but well) is considered to initiate a new discourse unit. However, while *mais* mostly functions as an initial marker, *mais bon* is more commonly used as a final attitudinal stance marker (see Hancil, 2015, for a similar observation on the English final but) .

(18) **Discourse Marker, final - initial confusion**
[et puis en fait non j' étais]$_{du}$ [$$] [bon] !! [j' avais pas trop le moral]$_{du}$
[and then in fact no I was]$_{du}$ [$$] [well] !! [I wasn't feeling too good]$_{du}$

(19) **Discourse Marker, final - initial confusion**
[ça s' est jamais bien passé je crois d' ailleurs] !! [mais bon]$_{du}$ # [là vraiment j' étais verte]$_{du}$
[It never went well I think by the way] !! [but well]$_{du}$ # [At that moment I was really disappointed]$_{du}$

Conversational spontaneous speech also present large bundles or discourse markers (Muller and Prévot, 2003). The model struggle to identify where the boundary should be, if any is really present in these sequences like in (20) in which the *mais (but)* is wrongly considered to segment the unit.

(20) **Discourse Marker, cluster**
[ah ben ouais] !! [mais ça]$_{du}$ # [non elles auraient dû au lieu d' emmener du des croissants]$_{du}$
[oh well yeah] !! [but this]$_{du}$ # [non they should have instead of bringing some some croissants]$_{du}$

Some positive feedback markers, in DU-initial position are also used as 'pivot' (Gravano et al., 2011) and in that case tend to be included in their DU-host like in (21), that the model failed to adapt to.

(21) **Discourse Marker, pivot**
[tu les as jamais trouvés ouais]$_{du}$ # [ouais #] !! [c' était hallucinant]$_{du}$
[you never found them yeah]$_{du}$ # [yeah #] !! [it was amazing]$_{du}$

5.3 The case of parentheticals

We would like to emphasize a last family of errors related to parentheticals (Asher, 2000). It was not possible to annotate truly embedded discourse units within another discourse units in our annotation model and framework.[24] While deliberate, this choice had some negative consequences for examples like (22) and more crucially (23) for the human coder decided to promote the embedded material as full-fledged discourse units. In a more expressive framework, these two examples would have been handled homogeneously, leading to a clearer situation for these parentheticals (even though they probably would have been harder to learn for the model).

(22) **Parentheticals**
[elles auraient dû prendre euh] !! [je sais pas une ou deux bières]$_{du}$
[they should have brought uh] !! [I don't know one or two beers]$_{du}$

(23) **Embedded Parenthetical**
[alors que y a]$_{du}$ [$$] [enfin euh à mon sens]$_{du}$ [$$] [y avait pas de norme]$_{du}$
[while there is]$_{du}$ [$$] [well uh in my sense]$_{du}$ [$$] [there was no norm]$_{du}$

5.4 Other sources of errors

The majority of the remaining errors relate to various spontaneous speech structures or specific uses, like some relatives that introduce clauses that are missed by the model like the first false negative in (24).

24. This choice was made for efficiency reasons after some pilot annotation studies with and without embedded units. Allowing embedded units made the whole process slower and more cumbersome while most of the units obtained this way were rather trivial and could be considered as stance markers for their host, in a medial position.

(24) **Relative**
[j'ai trouvé derrière le le micro-ondes] !! [donc une multiprise]$_{du}$ [$$] [où y avait à la fois le frigo]$_{du}$ [$$] [enfin # tout ce qui pouvait être branché su- quelque part]$_{du}$
[I found behind the the microwave] !! [so a power strip]$_{du}$ [$$] [where there was at the same time the fridge]$_{du}$ [$$] [well # everything that could be plugged on- somewhere]$_{du}$

As mentioned in the quantitative overview, non-sentential units (Fernández and Ginzburg, 2002; Fernández et al., 2007) were also a source of confusion like for (25) but also some versions of short canonical utterances (for the spontaneous speech genre) could pose problem like (26).

(25) **Non Sentential Units**
[des gens qui avaient des expériences]$_{du}$ [$$] [mais dans un domaine différent]$_{du}$
[people who had experiences]$_{du}$ [$$] [but in a different domain]$_{du}$

(26) **Spoken Canonical, very short utterance**
[bon tu commences]$_{du}$ [$$] [tu en as en tête]$_{du}$
[well you start]$_{du}$ [$$] [you have some in mind]$_{du}$

We can also mention rare but highly problematic cases of reported speech with dialogues present in the reported speech like (27).

(27) **Reported Speech, Dialogical**
[euh donc on a tout # effacé]$_{du}$ # [$$] [quoi]$_{du}$ # [mais pourquoi]$_{du}$
[um so we erased everything]$_{du}$ # [$$] [what]$_{du}$ # [but why]$_{du}$

5.5 Error Analysis Summary

To summarize the error analysis, the remaining errors produced by our models correspond to the expected categories. Fine-tuning the base model significantly improves performance by adapting it to *disfluencies, pauses, and discourse markers*. However, when these phenomena diverge further from the written pre-training data, because they involve combination of these phenomena, they are still misclassified by the model. It is important to note that some of these difficult cases are inherently ambiguous and require a deep understanding of discourse sequences to be resolved correctly. As mentioned in the error analysis, even human coders struggled to find a coherent segmentation when *disfluencies, pauses, and discourse markers* are intertwined into hard-to-interpret sequences.

The way pauses are handled, often over-trusted by the model, appears to be the most promising area for further improvement. Prosody will play a key role in further characterizing what happens before and after a pause, potentially allowing us to use pause duration in a more nuanced way.

Finally, the current approach, where the discourse flow is treated on a speaker-by-speaker basis rather than in a truly dialogical context, does not seem to cause significant errors compared to the simplicity gained in terms of pipeline and model. In fact, errors related to dialogical factors

account for only a small percentage of the total errors (see Appendix 11 for detailed numbers). Furthermore, some of these errors were actually related to dialogue occurring within the reported speech of a single speaker. Therefore, we do not plan to change the overall pipeline we adopted for this dataset. However, more interactional, heavily dialogical genres might benefit from a truly dialogical discourse segmentation approach.

6. Discussion

The previous sections have explored methods for performing automatic segmentation of *Elementary Discourse Units* (EDU) in spontaneous speech transcripts, achieving near-human-level results. This task is crucial in the current quantitative linguistic landscape. Indeed, without discourse segmentation, the alternatives for obtaining "sentence" or "utterance" units are limited to human segmentation, acoustic proxies (such as inter-pausal units), or syntactic proxies (such as re-punctuation methods). Our experiments show that, at least for our dataset, a relatively small amount of annotated data is sufficient to approach human performance. In contrast, the two proxies mentioned can only be considered as baselines.

Some might argue that our dataset is quite specific, combining dialogical and monological dimensions, making the proxies inefficient for part of the data. However, we consider on the contrary that most conversational situations combine or alternate between highly interactive sequences (short turns) and more expository, longer-turn segments (it's the case in meetings for instance). Thus, the proxies mentioned above are not sufficient.

We believe that the ability to segment large spontaneous speech corpora with these units opens up various possibilities for revisiting critical research questions, such as turn-taking, discourse prosody, multi-modal interfaces, and, more broadly, language structure as revealed by spontaneous speech. Having discourse units that are determined not only from syntax but also from broader semantic and discourse aspects is a step toward more efficient operationalization of other linguistic models. For instance, our discourse boundaries can be used as Transition-Relevant Places (TRPs) in the Conversational Analysis framework (Sacks et al., 1974), contributing to a more practical definition of conversational units, as suggested by Ford and Thompson (1996). They can also be systematically explored within dualistic frameworks (Haselow, 2017; Pietrandrea and Kahane, 2019).

Regarding turn-taking, using Inter-Pausal Units (IPUs) directly introduces circularity. Elementary Discourse Unit boundaries, however, provide an operational determination of the TRP, opening up the possibility of more accurate (than previous proxies) and large-scale corpus studies on turn-taking and turn transitions.

Together, with prosody extraction tools, discourse segmentation models of the kind presented here, can label with a near-human level large speech corpora. This is an important step for discourse-prosody interface research as it tend to be a highly variable phenomena and models would benefit from empirical evidence from larger corpora coming from more diverse sub-genres to boost their empirical foundation.

Concerning language structures, since pauses are not linguistic in nature, they are the source of untypical segmentations, leading to artificially complex structures to describe. The main reason is that pauses can be inserted almost anywhere. This is what happens in many disfluent or interaction sequences in which even rather long pauses ($>$ 500ms) do not interrupt discourse units. 'Sentences', on the other hand, reliably allow for identifying large language processing structures. Spontaneous spoken sequences thus differ radically through many different syntactic patterns from canonical traditional grammar as explained in (Ginzburg and Poesio, 2016).

There are several directions for further development of the current approach. Although we are near human-level performance, we are not there yet. As seen in the error analysis, many of the errors are tied to issues such as EDU boundaries without pauses or over-interpretation of pauses by the model. The next step is to model the prosodic context for each segmentation decision, with the aim of identifying the locations of prosodic breaks without pauses and distinguishing between pauses that do not correspond to real prosodic breaks. Another promising direction is to apply our model (perhaps with further fine-tuning on small-scale gold data) to other genres, corpora, and languages. As demonstrated in recent work, fine-tuning ROBERTA for lower-resource languages has proven to be a promising solution. Specifically, Prévot and Wang (2024) showed that fine-tuning ROBERTA yielded good results for Taiwan Southern Min, even though this language is not included in the pretraining corpus.

Acknowledgments

This work was conducted over several years and benefited from multiple sources of support. We gratefully acknowledge funding from the ANR project SUMM-RE (ANR-20-CE23-0017), the Institut Convergence ILCB (ANR-16-CONV-0002), and the Institut Universitaire de France (IUF). Laurent Prévot further thanks the Institute of Linguistics, Academia Sinica, for hosting him during the final stages of this research.

References

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. Automatic dialog act segmentation and classification in multiparty meetings. In *Proc. ICASSP*, volume 1, pages I–1061. IEEE, 2005.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Nicholas Asher. Truth conditional discourse semantics for parentheticals. *Journal of Semantics*, 17 (1):31–50, 2000.

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

John Langshaw Austin. *How to do things with words*. Harvard university press, 1975.

Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. Weak supervision for learning discourse structure. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2296–2305, Hong Kong, China, November 2019. Association for

Computational Linguistics. doi: 10.18653/v1/D19-1234. URL https://aclanthology.org/D19-1234.

Christophe Benzitoun and Frédéric Sabio. Où finit la phrase? où commence le texte?. l'exemple des regroupements de constructions verbales. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (7), 2010.

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

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

Claire Blanche-Benveniste, Mireille Bilger, Christine Rouget, Karel. Van Den Eynde, Piet Mertens, and Dominique Willems. Le français parlé (études grammaticales). *Sciences du langage*, 1990.

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

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

David Brazil. *A grammar of speech*. Oxford University Press, USA, 1995.

Harry Bunt. Multifunctionality in dialogue. *Computer Speech & Language*, 25(2):222–245, 2011.

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

Elizabeth Couper-Kuhlen and Margret Selting. Introducing interactional linguistics. *Studies in Interactional Linguistics. Amsterdam: John Benjamins*, pages 1–22, 2001.

Ludivine Crible and Liesbeth Degand. Domains and functions: A two-dimensional account of discourse markers. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24), 2019.

Viet-Trung Dang, Tianyu Zhao, Sei Ueno, Hirofumi Inaguma, and Tatsuya Kawahara. End-to-end speech-to-dialog-act recognition. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 3910–3914. ISCA, 2020. doi: 10.21437/Interspeech.2020-1062. URL https://doi.org/10.21437/Interspeech.2020-1062.

Liesbeth Degand and Anne Catherine Simon. Minimal discourse units: Can we define them, and why should we. *Proceedings of SEM-05. Connectors, discourse framing and discourse structure: from corpus-based and experimental analyses to discourse theories, Biarritz*, pages 14–15, 2005.

Liesbeth Degand and Anne Catherine Simon. On identifying basic discourse units in speech: theoretical and empirical issues. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (4), 2009.

Henri-José Deulofeu. L'approche macrosyntaxique en syntaxe: un nouveau modèle de rasoir d'occam contre les notions inutiles? *Scolia: Sciences Cognitives, Linguistiques et Intelligence Artificielle*, 16(1):77–95, 2003.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. The llama 3 herd of models. *arXiv preprint 2407.21783*, 2024. URL `https://arxiv.org/abs/2407.21783`.

Benoit Favre, Dilek Hakkani-Tur, Slav Petrov, and Dan Klein. Efficient sentence segmentation using syntactic features. In *2008 IEEE Spoken Language Technology Workshop*, pages 77–80. IEEE, 2008.

Raquel Fernández and Jonathan Ginzburg. Non-sentential utterances: Grammar and dialogue dynamics in corpus annotation. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics, 2002.

Raquel Fernández, Jonathan Ginzburg, and Shalom Lappin. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427, 2007.

Cecilia E Ford and Sandra A Thompson. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Studies in interactional sociolinguistics*, 13:134–184, 1996.

Chris Fournier and Diana Inkpen. Segmentation similarity and agreement. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 152–161, Montréal, Canada, 2012.

Kim Gerdes and Sylvain Kahane. Speaking in piles: Paradigmatic annotation of french spoken corpus. In *Fifth Corpus Linguistics Conference*, pages 1–15, 2009.

Jonathan Ginzburg. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, 2012.

Jonathan Ginzburg and Massimo Poesio. Grammar is a system that characterizes talk in interaction. *Frontiers in Psychology*, 7, 2016.

John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE, 1992.

Agustín Gravano, Julia Hirschberg, and Štefan Beňuš. Affirmative cue words in task-oriented dialogue. *Computational Linguistics*, 38(1):1–39, 2011.

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

Oliver Guhr, Anne-Kathrin Schumann, Frank Bahrmann, and Hans Joachim Böhme. Fullstop: Multilingual deep models for punctuation prediction. June 2021. URL http://ceur-ws.org/Vol-2957/sepp_paper4.pdf.

Narendra K Gupta and Srinivas Bangalore. Segmenting spoken language utterances into clauses for semantic classification. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721)*, pages 525–530. IEEE, 2003.

Sylvie Hancil. The grammaticalization of final but: From conjunction to final particle. *Final particles*, pages 197–218, 2015.

Alexander Haselow. *Spontaneous spoken English: An integrated approach to the emergent grammar of speech*. Cambridge University Press, 2017.

Marti A Hearst. Multi-paragraph segmentation expository text. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, 1994.

Sophie Herment, Laetitia Leonarduzzi, Diana Lewis, Cristel Portes, Laurent Prévot, Frédéric Sabio, and Gabor Turcsan. Périphéries gauche et droite. *TIPA. Travaux interdisciplinaires sur la parole et le langage*, (38), 2022.

Julia Hirschberg and Barbara Grosz. Intonational features of local and global discourse structure. In *Proceedings of the DARPA workshop on Spoken Language Systems*. Association for Computational Linguistics, 1992.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted JM Sanders. Segmenting discourse: Incorporating interpretation into segmentation? *Corpus Linguistics and Linguistic Theory*, 14(2):357–386, 2018.

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

Dan Jurafsky, Liz Shriberg, and Debra Biasca. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. Technical report, University of Colorado at Boulder, 1997.

Diana M Lewis. Pragmatic markers at the periphery and discourse prominence. *Pragmatic markers and peripheries*, 325:351, 2021.

Per Linell. *The written language bias in linguistics: Its nature, origins and transformations*. Routledge, 2004.

Pierre Lison, Jeremy Barnes, and Aliaksandr Hubin. skweak: Weak supervision made easy for NLP. In Heng Ji, Jong C. Park, and Rui Xia, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 337–346, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.40. URL https://aclanthology.org/2021.acl-demo.40/.

Yang Liu, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. Using conditional random fields for sentence boundary detection in speech. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05)*, pages 451–458, 2005.

William C Mann and Sandra A Thompson. Rhetorical structure theory: Description and construction of text structures. In *Natural language generation: New results in artificial intelligence, psychology and linguistics*, pages 85–95. Springer, 1987.

Daniel Marcu. The theory and practice of discourse parsing and summarization, 2000.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41 (3):437–479, 2015.

Eleni Metheniti, Chloé Braud, Philippe Muller, and Laura Rivière. 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, 2023.

Eleni Metheniti, Philippe Muller, Chloé Braud, and Margarita Hernández Casas. Zero-shot learning for multilingual discourse relation classification. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17858–17876, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1553.

Philippe Muller and Laurent Prévot. An empirical study of acknowledgment structures. In *7th workshop on the semantics and pragmatics of dialogue*, 2003.

Philippe Muller, Chloé Braud, and Mathieu Morey. 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, 2019.

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. Are LLMs better than reported? detecting label errors and mitigating their effect on model performance. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26770–26797, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1360. URL https://aclanthology.org/2025.emnlp-main.1360/.

Kota Shamanth Ramanath Nayak. Does ChatGPT measure up to discourse unit segmentation? a comparative analysis utilizing zero-shot custom prompts. 2024.

Joakim Nivre. Dependency parsing. *Language and Linguistics Compass*, 4(3):138–152, 2010.

Berthille Pallaud, Stéphane Rauzy, and Philippe Blache. Auto-interruptions et disfluences en français parlé dans quatre corpus du cid. *TIPA. Travaux interdisciplinaires sur la parole et le langage*, (29), 2013.

Rebecca J Passonneau and Diane Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139, 1997.

Klim Peshkov and Laurent Prévot. Segmentation evaluation metrics, a comparison grounded on prosodic and discourse units. In *LREC*, pages 321–325, 2014.

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

Lev Pevzner and Marti A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

Janet Pierrehumbert and Julia Bell Hirschberg. The meaning of intonational contours in the interpretation of discourse. In *Intentions in communication*. MIT press, 1990.

Paola Pietrandrea and Sylvain Kahane. Macrosyntactic annotation. *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French, John Benjamins, Amsterdam*, pages 97–126, 2019.

Livia Polanyi and Remco JH Scha. The syntax of discourse. *Text-Interdisciplinary Journal for the Study of Discourse*, 3(3):261–270, 1983.

Livia Polanyi, Chris Culy, Martin Van Den Berg, Gian Lorenzo Thione, and David Ahn. A rule based approach to discourse parsing. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 108–117, 2004.

Salvador Pons Bordería. Models of discourse segmentation in romance languages: An overview. *Discourse Segmentation in Romance Languages*, pages 1–21, 2014.

Cristel Portes and Roxane Bertrand. Permanence et variation des unités prosodiques dans le discours et l'interaction. *Journal of French Language Studies*, 21(1), 2011.

Laurent Prévot and Sheng-Fu Wang. 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, 2024.

Laurent Prévot, Roxane Bertrand, Klim Peshkov, Stéphane Rauzy, and Philippe Blache. Prosody, discourse and syntax in french conversations: Resource creation and evaluation. *(submitted)*, 2016.

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

Laurent Prevot, Julie Hunter, and Philippe Muller. Comparing methods for segmenting elementary discourse units in a french conversational corpus. In *24th Nordic Conference on Computational Linguistics (NoDaLiDa 2023)*. ACL: Association for Computational Linguistics; University of Tartu Library, 2023.

Laurent Prévot, Roxane Bertrand, and Julie Hunter. Segmenting a large french meeting corpus into elementary discourse units. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 183–191, 2025.

Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman. jiant: A software toolkit for research on general-purpose text understanding models. In Asli Celikyilmaz and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 109–117, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.15. URL https://aclanthology.org/2020.acl-demos.15/.

Silvia Quarteroni, Alexei V. Ivanov, and Giuseppe Riccardi. Simultaneous dialog act segmentation and classification from human-human spoken conversations. In *Proc. ICASSP*, pages 5596–5599, 2011. doi: 10.1109/ICASSP.2011.5947628.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 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, 2017.

Eddy Roulet, Filliettaz Laurent, Anne Grobet, and Marcel Burger. *Un modèle et un instrument d'analyse de l'organisation du discours*. Bern: Lang, 2001.

Frédéric Sabio. Phrases et constructions verbales: quelques remarques sur les unités syntaxiques dans le français parlé. In *Constructions verbales et production de sens*, pages 127–139. Presses Universitaires de Franche-Comté, 2006.

Harvey Sacks. *Lecture on conversations*. Basil Blackwell, 1992.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *language*, pages 696–735, 1974.

Emmanuel A. Schegloff and Harvey Sacks. Opening up closings. *Semiotica*, 8, 1973.

Margret Selting. The construction of units in conversational talk. *Language in society*, 29(4): 477–517, 2000.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication*, 32(1):127–154, 2000.

John Sinclair and Malcolm Coulthard. Toward an analysis of discourse. In *Advances in spoken discourse analysis*. Routledge, 1992.

Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235, 2003.

Manfred Stede. *Discourse processing*, volume 15. Morgan & Claypool Publishers, 2012.

Amanda Stent. Rhetorical structure in dialog. In *INLG'2000 proceedings of the first international conference on natural language generation*, pages 247–252, 2000.

Andreas Stolcke and Elizabeth Shriberg. Automatic linguistic segmentation of conversational speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 2, pages 1005–1008. IEEE, 1996.

Andreas Stolcke, Elizabeth Shriberg, Rebecca A Bates, Mari Ostendorf, Dilek Zeynep Hakkani, Madelaine Plauche, Gökhan Tür, and Yu Lu. Automatic detection of sentence boundaries and disfluencies based on recognized words. In *ICSLP*, volume 2, pages 2247–2250. Citeseer, 1998.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26 (3):339–373, 2000.

Zeno Vendler. Verbs and times. *The philosophical review*, pages 143–160, 1957.

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 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, 2019.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene, editors. *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.disrpt-1.0`.

Tianyu Zhao and Tatsuya Kawahara. A unified neural architecture for joint dialog act segmentation and recognition in spoken dialog system. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–208, 2018.

Appendix A. Additional Inter-coder agreement

Spk	AB	AC	AG	AP	BX	CM	EB	IM	LJ	LL	MB	MG	ML	NH	SR	YM	mean
Annot	A,E	B,D	D,A	C,E	A,E	A,E	C,E	D,A	C,E	D,C	B,D	A,E	D,A	D,C	C,E	D,A	
κ	0.854	0.857	0.829	0.839	0.868	0.846	0.856	0.825	0.840	0.868	0.853	0.841	0.860	0.856	0.848	0.823	0.848

Table 8: κ scores for the discourse boundaries.

Spk	AG	LL	NH	YM	mean
Annot	A,D,B	A,C,D,E	A,C,D,E	A,D,B	
κ	0.837	0.783	0.842	0.853	0.829

Table 9: Multi-κ scores for the discourse boundaries on 15 minute fragments by 4 speakers, annotated by 3 or 4 naive annotators.

Annotator	Expert 1	Expert 2	mean
Annot A	0.753	0.803	0.778
Annot C	0.809	0.877	0.843
Annot D	0.811	0.885	0.848
Annot E	0.805	0.880	0.843
mean			0.828

Table 10: κ scores for the discourse boundaries of each naive annotator versus expert on 15 minute fragments by two (30 minutes in total).

Inter-annotator agreement on discourse units between naive annotators is reported in Tables 8 and 9. The whole duration for each speaker was annotated by two naive annotators, the κ scores per speaker are shown in table 8. Table 9 contains multi-κ values for 15 minute excerpts of 4 speakers, two of these extracts were annotated by 3 annotators and the other two by 4 annotators. According to these scores, the inter-annotator agreement for discourse boundaries is consistently high across speakers and annotators.

Appendix B. Supplementary Figures and Tables

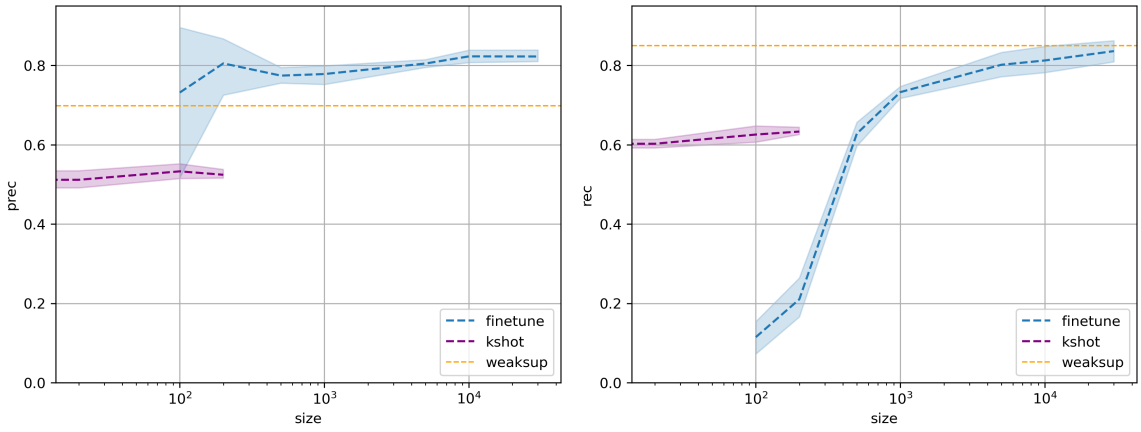


Figure 4: Precision / Recall comparing fine-tuning and few-shots approaches. x-axis: nb of tokens, log scale. Shaded areas are 95% confidence intervals.

Main Apparent Error Cause	Count
disfluencies	60
discourse marker	48
spoken structure	28
non sentential unit	14
dialogical	7
canonical	7
only pause	5
reported speech / attribution	5
relative clause	2

Table 11: Error Analysis Detailed Numbers

Appendix C. Examples of Labelling Functions

Pseudo-code in python. Each document indexes its tokens. Tokens have a few attributes, mostly their forms and part-of-speech tags, and a duration in case of a pause. Rules output a label, and a token index position, coding if a boundary label must be before or after the token (as this is rule dependent).

```
NON_END_TOK = PRO_SUJ + PRO_OTH + PRO_REL + DEM + DET + PREP + OTH
              + NEG + DM_INI + RELATIVES

def non_end_tok(doc):
    for idx, token in enumerate(doc):
```

```
        if idx > 1:
            if doc[idx - 1].text in NON_END_TOK:
                yield idx, idx + 1, NOBEG


def long_pause(doc):
    for idx, token in enumerate(doc):
        if idx > 0:
            if doc[idx - 1].text in ['#'] and doc[idx - 1]._.dur > 0.8:
                if doc[idx].text not in ['#']:
                    yield idx, idx + 1, BEG
                    yield idx - 1, idx, NOBEG
                else:
                    if idx < len(doc) - 1:
                        yield idx + 1, idx + 2, BEG
                    else:
                        yield idx, idx + 1, BEG
        else:
            if doc[idx].text not in ['#']:
                yield idx, idx + 1, BEG
```

Appendix D. Few-shot Prompt example

```
Your task is to segment a dialog in elementary discourse units,
where the character # indicates a short pause

Here are a few examples of the expected results:

Example 1:
input:  ah mais c' est bon je sais plus quoi dire là c' est bon
hein i- t- # enfin je sais pas
output:  | ah mais c' est bon | je sais plus quoi dire là | c' est
bon hein | i- t- # enfin je sais pas

Example 2:
input:  ah ouais ah ouais et donc si tu veux il avait acheté
tu sais des bouquins de euh # je me demande si c' était pas le
Reboul là tu sais le g- le le # la bible de la cuisine provençale
enfin je sais plus # et bref il voulait faire d- # de la crème de
marrons # tu vois # donc euh # i- il prend la * recette et tout
bon il v- il dit bon ok # ce c- # cet idiot tu sais ce qu' il fait
# il va prendre des marrons tu t' en rappelles # il va prendre des
marrons sur le cours Mirabeau ou dans le parc Jourdan je sais pas
quoi enfin bref il ramasse ses marrons pourris quoi et # il les
pèle # enfin je sais pas à quoi il pense il a pris une journée à
```

faire la crème de marrons
output: | ah ouais | ah ouais | et donc si tu veux il avait acheté tu sais des bouquins de euh # | je me demande si c' était pas le Reboul là | tu sais le g- le le # la bible de la cuisine provençale | enfin je sais plus # | et bref il voulait faire d- # de la crème de marrons # tu vois # | donc euh # i- il prend la * recette et tout bon | il v- il dit | bon ok # | ce c- # cet idiot tu sais ce qu' il fait # | il va prendre des marrons | tu t' en rappelles # | il va prendre des marrons sur le cours Mirabeau ou dans le parc Jourdan je sais pas quoi | enfin bref | il ramasse ses marrons pourris quoi | et # il les pèle # | enfin je sais pas à quoi il pense | il a pris une journée à faire la crème de marrons |

Example 3
input: ça c' était les # les chocs culturels
output: | ça c' était les # les chocs culturels # |

Example 4
input: et ouais en plus elle toute seule du coup # bon là c' est pas rigolo parce qu' ils vont déménager mais # mais au moins i- il sera là-haut quoi tu vois
output: | et ouais en plus elle toute seule du coup # | bon là c' est pas rigolo | parce qu' ils vont déménager | mais # mais au moins i- il sera là-haut quoi tu vois

Exemple 5
input: ouais ouais # ou en vrai aussi des fois non # ah ouais
output: | ouais ouais # | ou en vrai aussi des fois non # | ah ouais |

Here is the text to segment:

on l' assimile # après on va la di-

Only print the original text, indicating segment boundaries with a | character, and do not add anything, do not remove anything. Do not present the result with introductory text.