Prosody, Discourse and Syntax in French Conversations: Resource creation and Evaluation

Laurent Prevot, Roxane Bertrand, Klim Peshkov, Stéphane Rauzy, Philippe Blache

Unpublished Manuscript manuscript No.
(will be inserted by the editor)

Prosody, Discourse and Syntax in French Conversations: Resource Creation and Evaluation

Laurent Prévot · Roxane Bertrand · Klim Peshkov · Stéphane Rauzy · Philippe Blache

January 2017

Abstract How syntax, prosody and discourse do interface with each other is a recurrent question in modern linguistics. While strong inter-dependencies have been established for read speech, the question is more challenging in the context of truly spontaneous conversational speech. To our best knowledge, there is not yet a reliable resource for addressing this question from a quantitative viewpoint, at least for French language. We present in this paper the process of creation of such a resource. We started from the Corpus of Interactional Data (CID) [13] and organized an annotation campaign for marking prosodic phrasing, discourse segmentation and disfluencies tagging. Syntactic tagging and chunking have been produced automatically and the former had been manually corrected. We describe here the methodology, the challenges encountered and the results of the annotation campaign. Moreover, we evaluate the campaign and propose a reference dataset from the annotations produced. All the resources described are available trough Ortolang perennial language archive repository.

Keywords Conversation · Discourse · Prosody · Syntax · Evaluation · Linguistic Units

1 Introduction

Studying of the interfaces between linguistic domains became a major trend in the last two decades. Descriptive, computational and experimental accounts have been developed. In the domain of discourse, prosody and syntax, these studies either remained descriptive [38,86,40,14] or limited to laboratory examples as in [70] for prosody-syntax and [41] for prosody-discourse. This situation is at least partly due to the difficulty to build 'real-life' datasets and systematically annotate them with the linguistic domains involved. Moreover in most studies, there is a more or less overtly admitted bias toward one of the domain concerned. We need a reference corpus for the study of these interfaces, specially for French language, a widely spoken language but dramatically less represented language than English with regards to language resource production. The work presented here is an attempt to produce a resource for studying the prosody/syntax/discourse interplay in a conversational setting, a genre in which speech is fiercely spontaneous. This work is perfomed on the Corpus of Interactional Data (CID) [13] and the annotations were performed in the framework of the ToMA (Tools for Multimodal Annotation) [18].[1]

Aix Marseille Université and CNRS
5 Av. Pasteur
13100 Aix-En-Provence
France
Tel.: +33-413-553-596
E-mail: firstname.lastname@lpl-aix.fr

[1] In French, Outils pour le Traitement de l'Information Multimodale (OTIM).

As stated before, previous descriptive work has been made on French language. There are however the results of individual expert researchers that did not systematically evaluate their analyses. Our goal here was different: experts established annotation guidelines that were given to semi-naive annotators. We then evaluate their agreement with different measures and discussed the possibility of establishing gold standards serving as reference for conversational data. The recent outcome of the RHAPSODIE project [65] is an interesting complementary resource to the one described here. However, RHAPSODIE resource is a compilation of different speaking styles and conversational situation while what we wanted here is a larger dataset for conversational speech. Moreover the principles underlying the discourse and even more the prosodic segmentation are different for the two projects, as we will see in sections 3 and 4.

Our objectives are (i) to present in details the annotation of a few domains of the larger project ; (ii) to constitute a first step toward a reference corpus for prosodic phrasing / discourse units of conversational speech ; and (iii) to present some insights gained about the annotation process for producing such a corpus. We start (Section 2) by proposing a brief overview of the project within which the resources described have been produced. Then we address two specific large-scale annotation experiments: Prosodic phrasing (Section 3) and Discourse Units segmentation (Section 4). These sections summarize the annotation guidelines, some examples to illustrate the data produced and discuss the annotation task itself. The question of syntax is addressed (in Section 5) from a slightly different perspective since this level had been first automatically produced and only then manually corrected. We also provide a complete discussion of the evaluation in section 6. Before concluding, we present a quantitative and qualitative overview (Section 7) of the resource produced and discuss how this overview related to previous work on these interfaces.

2 The ToMA Project and related work

Several projects addressed the question of multimodal resources and their annotation but the ToMA project [18] presents its specificities and is unique in its content, at least for French language. The LUNA project [93,94] focuses on spoken language understanding. Its corpus is made of human-machine and human-human dialogues and proposes, on top of the transcription, different levels of annotation, from morpho-syntax to semantics and discourse analysis. SAMMIE [62] is another project aiming at building multimodal resources in the context of human-machine interaction. Annotations are done using the Nite XML Toolkit [25]; they concern syntactic and discourse-level information, plus indication about the specific modality used in the experiment. Aside, these work focusing on human-computer dialogues, important resources have been built, for English, for human-human conversations either for speech technologies training purposes with Switchboard corpus [51] or for linguistic research [3] with HCRC Map Task Corpus. Switchboard is a huge corpus but only features rather short phone conversations between strangers and therefore cannot exhibit the same phenomena found in more realistic conversational setting. HCRC MapTask is an amazing corpus that had been investigated in all directions but setting is strongly task-based and therefore present clear differences with conversational speech. As explained in the introduction, Rhapsodie project [65] is an interesting complementary resource that gathers various communicative styles and in which therefore, conversational speech is only seldom represented. Moreover the theoretical principles underlying prosodic, discourse and syntax enrichments are rather different from the work presented here as we will see in the next sections. The on-going Orfeo project[2] extends quantitatively Rhapsodie's work but still does not focus on conversational speech and employs mostly automatic techniques to perform the annotations. Finally, in terms of contents the closest resource is the The Nijmegen Corpus of Casual French [101] which presents similar data to ours but only transcription without further linguistic annotation. Overall, the Corpus of Interactional Data, together with the annotation described in this paper, presents a unique dataset, a least for French language, of relatively large amount of conversational speech annotated with various linguistic domains.

[2] Orfeo project website: `http://www.projet-orfeo.fr/`

After this overview of the project, its relations with other resources and a general presentation of its specificities, we will focus in the following sections on prosody, discourse and syntax domains.

Annotating corpora first requires to specify what kind of information it is necessary to represent and how it is organized. This problem consists in defining a coding scheme. Several of them have been developed in different projects. What comes out is that they are very precise in one or two modalities. However, they usually do not cover the entire multimodal domain nor the very fine-grained level of annotation required in every modality. We propose to combine several existing schemes and to extend them so as to obtain a coding scheme that would be as complete as possible.

A general coding scheme is of deep importance not only in terms of standardization and knowledge representation, but also for practical matters: it constitutes the basis for a pivot language, making it possible for the different tools (Praat, Anvil, etc.) to exchange formats. This is one of the goals of the PAULA format [28]. From the same perspective, starting from an adaptation of this format, we are developing tools implementing such interoperability, relying on a translation between the source format of the tool and this language. In parallel to the development of our corpus and annotation campaign, the effort toward the standardisation of linguistic resources [56] had been continued. The timing did not allow to frame our annotation into this framework but it is clear that our annotations can be ported to this annotation standard, for example within ISO-Diaml [22] for the discourse part. For more details on the OTIM project, see [18], for the CID corpus see [13] and for the general annotation schema, [17,19].

3 Prosodic Boundaries annotation

3.1 Annotation description

The prosodic level can be annotated in a manual or an automatic way depending on whether we observe rather phonological (more abstract) phenomena or phonetic parameters. In OTIM project, we focused on prosodic phrasing which corresponds to the structuring of speech material in terms of boundaries and groupings. Our global aim was to develop a phonologically-based transcription system for French that would be consistent enough to be amenable to automatic labeling. The first step was to compare manual annotations. This type of annotation is very time consuming but was necessary to improve the knowledge of prosodic domains in French. A second step will consist in improving existing automatic tools (such as Intsint for example [54]), by comparing the output of different annotation tools and manual expert/naive annotation [80,79].

The creation of the guidelines presented here is based on various annotations performed on the spontaneous data of the CID by experts during the last few years (see [14,13,72,85] among others). Thanks to these annotation campaigns, prosodic phrasing has been identified as the most urgent and the most relevant aspect to investigate for spontaneous speech. The main reasons are: (i) The segmentation in groups and boundaries is crucial for the interpretation process; (ii) This segmentation is essential at every linguistic level (it enables a comparison between the different units) ; and (iii) models of prosodic phrasing were mainly elaborated on very controlled data and it is crucial to show the sustainability of the phonological units in spontaneous data.

The different phonological models of the prosodic structure of French [39,32,33,60,59,87,68] have in common that French language is characterized by two levels of phrasing. The minor prosodic phrase or accentual phrase (henceforth `AP` or level 1)[3] [58,59] is the lowest boundary level in French. `AP` is associated with a final/primary accent on the final full syllable (non schwa) of the content word and an optional secondary accent associated with the beginning of the phrase. While the final accent is characterized by a melodic variation (more often a rise) and a syllabic lengthening,

[3] The different phrases received two names: `AP` being more theory-loaded than the simple level 1. This was important for our guidelines for semi-naive coders but in the discussions the two names are equivalents.

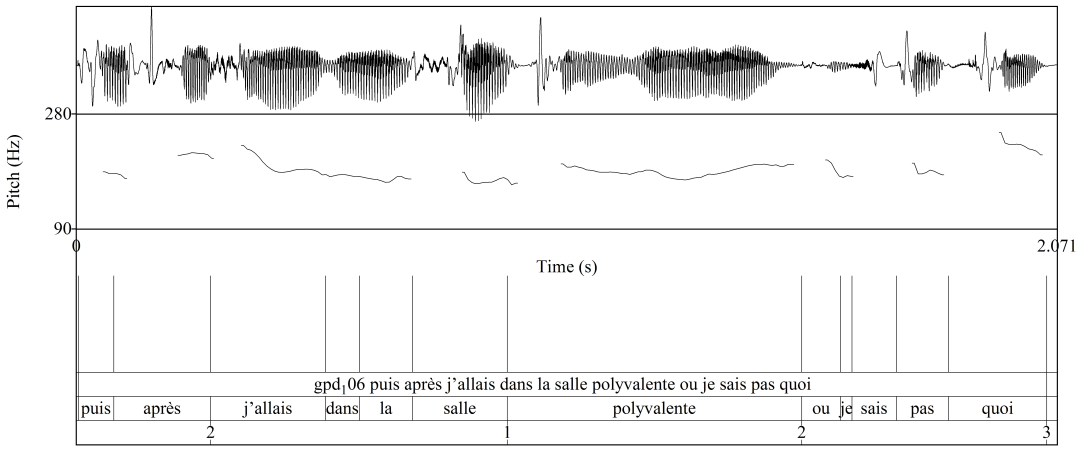

Fig. 1 AP, ip and IP coded as 1, 2 and 3 respectively

the secondary/initial accent is only characterized by a rise [76,7,106] (See Figure 1). Composed of one or more AP, the major unit or intonational phrase (`IP` or level 3) is the highest boundary in the hierarchy and the domain of the intonative contour. It is also marked by lengthening and melodic variations but more important than for `AP` (See Figure 1). Although the third level of phrasing is not yet well established in French (see however [59], [40], and more recently [69,42]), and though the definition still requires to be refined more specifically for spontaneous speech, we added this intermediate phrase (`ip` or level 2). `ip` would be the domain of the downstep (result of a global lowering of the pitch range level on this specific phrase, for details see [69,42]). Our aim here was not to give so precise details about phonetic correlates of each unit but only to test whether annotators needed a third level of unit in spontaneous data (is a third level perceptually relevant?), regardless its phonetic correlates. Therefore we only described `ip` as the intermediate unit between `AP` and `IP` (characterized by melodic and duration cues between `AP` and `IP` as well [69]). In this way, the `ip` would be present only when the stretch of discourse is sufficiently long for distinguishing between `ip` and `IP`. The three levels of boundaries will be thus coded in a relative way by the annotators (See Figure 1 again).

By focusing on a single aspect (prosodic structure) we leave out other several relevant prosodic phenomena in conversation such as non categorical phenomenon like pitch range or tempo. We think that it is more important to make the prosodic structures annotation available and to capture by some quantitative measure other types of parameters such as the duration of syllables inside each prosodic constituent. Annotators however were asked to pay attention to these gradient prosodic phenomena to annotate prosodic structure. Conversation is frequently marked by stretch of speech with global variations of tempo or pitch range [85].

Given that the different units have been established for laboratory speech, they first remain to be validated on spontaneous data (see [85] for a previous attempt to show that these categories remain relevant, for an adaptation to this type of data). Thanks to the first experiment and several annotation experiments from experts, results have shown a very good inter-coder agreement for the higher level of constituency (`IP`) (see [72] and section 6).

Guidelines for transcribing prosodic units in French by naive annotators have been elaborated in order to test the replicability of the annotation criteria. But more complete and well-informed guidelines for the different prosodic units listed in the literature have been established. Firstly, the previous annotation phase by experts using only these 2 levels of phrasing highlighted a missing

intermediate unit and worked toward an additional level. Secondly, the experts had to adapt to this type of data by introducing one more category at the level of IPs: an uncompleted IP (ipa) corresponding to a stretch of speech larger than AP which wasn't completed. More globally spontaneous data indeed exhibit different phenomena such as word repeats, strong syllabic lengthening, fillers (such as *'euh'/'uh'*) and silent pauses that can make the prosodic phrasing annotation more difficult.

Guidelines and annotation process The guidelines start by briefly presenting the prosodic system of French in order to obtain a more phonologically-based transcription than a transcription based on syntactic or acoustic criteria only. It is important for annotators to keep in mind that prosodic phrases are connected with phonetic properties that are associated with abstract phonological segments or features. The guidelines systematically provide examples to illustrate each point or phenomenon. They are provided as external resources of this paper.

Each conversation was blindly cross-segmented by at least two coders. After reading of the guidelines, a training step consisted of annotating a few short extracts of the corpus and included 3 meetings (for debriefing) with an expert. The corpus was annotated following a perceptually-based procedure. Prosodic boundaries can be marked differently: either words are grouped into prosodic phrases either each word is marked with a break index (degree of juncture) [23, p406]. In our study, we used the latter method (as in ToBI).

3.2 Break indices tier

Our guidelines are based on work by several authors in French [39,87,59,60] and also inspired from Tones and Break Indices (ToBI) guidelines [9] in which *the break index tier marks the prosodic grouping of the words in an utterance by labeling the end of each word for the subjective strength of its association with the next word, on a scale from 0 (for the strongest perceived conjoining) to 4 (for the most disjoint)*. The value of the different labels in French do not correspond to the value of the labels in English ToBI since the prosodic structure (number and size of constituents) in French and in English are different. While for English, 1 corresponds to the word, 2 and 3 to the intermediate phrase (2 when ip is only marked by the duration cue) and 4 to the intonational phrase, in the present work, the label 1 corresponds to the accentual phrase, 2 to the intermediate phrase, and 3 to the intonational phrase. 0 refers to the less disjuncture between 2 words and is automatically and a posteriori annotated. Another difference with English ToBI guidelines is that tones associated with pitch accents are not yet provided in French: English ToBI contains two separate markings relative to the boundaries of an intonational phrase (one tier for pitch accent + boundary tone and another tier for the break index). This is why our guidelines had to explain the prosodic system of French in terms of accentuation since initial and final accents are markers of the prosodic structure. It is however an additional difficulty for the annotation process. Finally, as already mentionned above, when disfluent segments prevented the identification of the prosodic phrasing, the annotators had to note *disf*. The annotation of phrasing then consists of four tiers: Inter Pausal Units[4] and tokens tiers initially provided as well as two empty tiers for break indices and disfluencies. Each annotations label was instructed to be anchored on token boundaries which is therefore the reference tier. The resulting picture of these instructions can be seen on tier 3 of figure 1.

Data available to the coders The segmentation was performed on time-aligned data from both participants with an access to the signal and f0 (pitch) curve. The segmentation was done with the tier-based tool Praat [21]. Even if annotators did not make a tonal transcription, they used the f0 curve to identify some melodic variations associated with final and initial accents.

[4] Inter Pausal Units (IPUs) are stretchs of continuous speech separated by 200 or more milliseconds silent pauses. IPUs are generally automatically created but because of some residual spill, some manual correction was necessary.

Specific traps described in the guidelines In the guidelines, we presented a few cases in which a bad interpretation of the fundamental frequency pattern was due to other aspects than prosody. For example, the most frequent and the most known effect is that consonant segment in an utterance interrupt the smooth course of the fundamental frequency. This perturbation of the f0 curve is not a cue of boundary. Another aspect concerns the presence of extra-metric syllable. The identification of relevant final melodic variations in French can be disturbed when the last syllable contains a *schwa*. In the example of figure 2, we can observe a rise on the penultimate syllable (*me*) and an edge tone associated with n@. It is a rare case in French where we can distinguish between pitch accent and boundary tone thanks to the presence of the extrametrical syllable (schwa). Finally, due to the fundamental role of the syllable lenghtening (prelenthtening boundary) or silent pause in the identification of a prosodic boundary, annotators had to handle disfluencies separately.

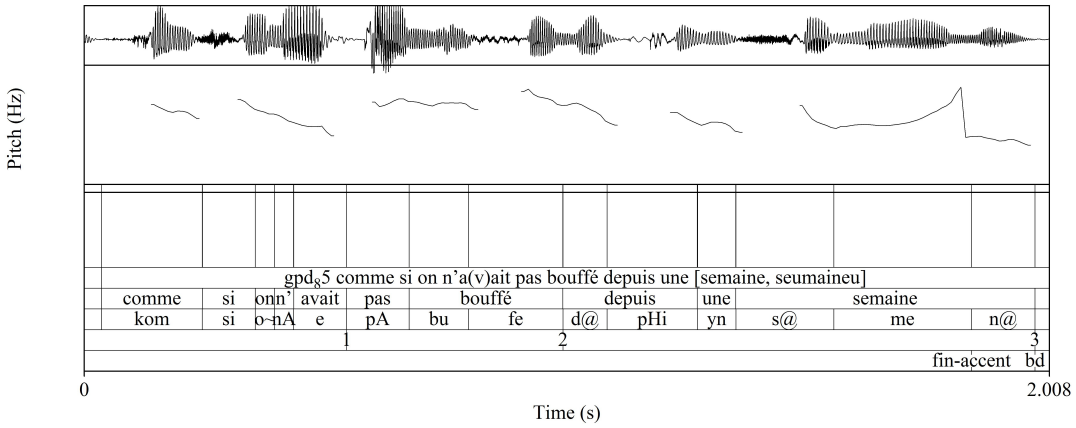

Fig. 2 A schwa syllable as a cue of boundary

3.3 Disfluencies tier

As explained in section 3.1, it was necessary to mark disfluencies in addition to the break indices. An independent tier was therefore created for annotating disfluencies as illustrated in Figure 3. In this figure, the end of the disfluent segment is annotated when the speech becomes fluent again. All these phenomena were then included in a specific category *disfluency*.

Syllabic lengthening While syllabic lengthening is a crucial cue for identifying accent and boundary (pre-lengthening boundary), it can also be a cue of disfluency. We asked annotators to distinguish, when it is possible, between a lengthening associated with a disfluency and a lengthening associated with an accent. In the example presented in figure 3, the strong lengthening on the determiner *'un' ('a')* just before the noun *'bistrot' ('bar')* (together with a strong creaky voice that blocks voicing) is associated with a disfluent mark but not with a boundary mark. It is important to note that a disfluency does not always prevent the identification of the prosodic structure [85].

The silent pauses Another parameter frequently involved in the identification of a boundary is the silent pause. Spontaneous data show that silent pauses (which inherently contribute to the impression of a break) do not necessarily imply a prosodic boundary. Prosodic phrasing can be altered by pauses associated with disfluent segments that do not induce a true major prosodic boundary. In this way, [85] recommended to distinguish between a pause as a disfluent cue and a real cue of boundary. Furthermore, prosodic phrasing can be preserved thanks to the intonative

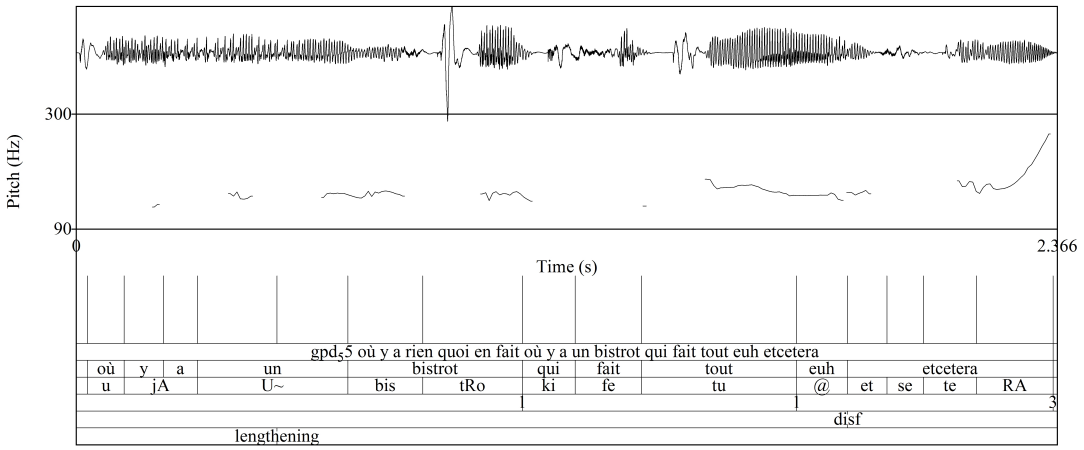

Fig. 3 Disfluent lengthening not associated with a boundary

cohesion (a form of continuity) between items around pauses when silent pauses function as a rhetoric or stylistic device [43].

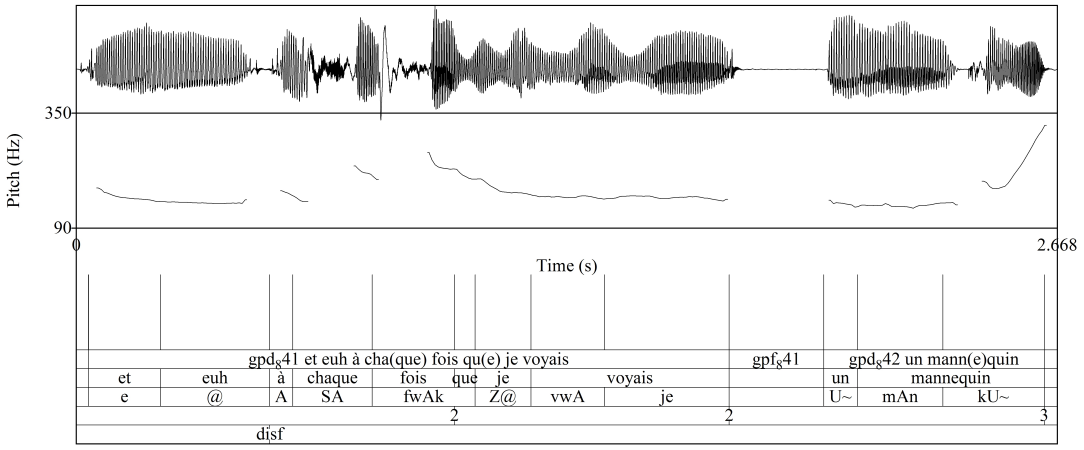

Fig. 4 Example of stylistic silent pause

In both cases (Figures 4 and 5), silent pauses do not interrupt the intonative cohesion of the major unit (marked 3). These silent pauses are stylistic and are used by the speaker to highlight the following words. In the second figure, the speaker seems to spell each item (separated by silent pause) highlighting it and without the presence of pause interrupts the major unit.

3.4 Deriving Prosodic Units for a gold standard

For the purposes of a comparative study [89], we needed prosodic units (`PU`) as intervals and not as boundaries. We created this annotation automatically based on the following rules:

– Only boundaries of levels 2 and 3 are considered relevant for our 'bigger' prosodic units datasets

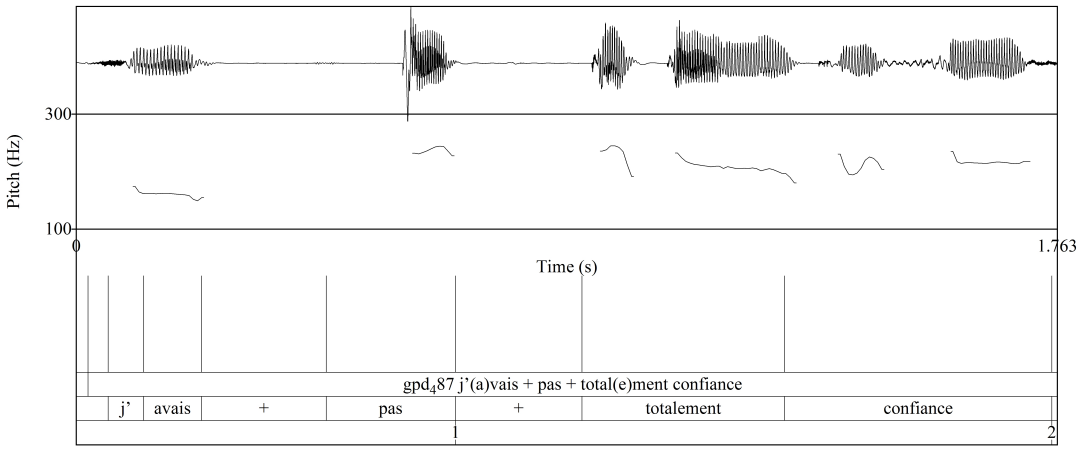

Fig. 5 Example of stylistic silent pause

- Although all the annotators have used the three levels of units, there was great variability of the level 2. Moreover for ensuring replicability of the annotation, boundaries of level 2 and 3 are finally merged.
- Any pause of 800ms is treated as a unit separation and therefore we used the pause left and right boundaries respectively as right boundary of the preceding unit and left boundary of the following unit.

The value of 800ms was used to be sure to include pauses that may not break the current prosodic unit. This was a strategy to correct many obvious breaks that were not marked by the naive annotations due to over-interpretation of the guidelines on this point. It also helps segmenting more systematically the backchannels for which no specific instructions were given. A comparison of the scores between raw prosodic boundaries annotations and our derived prosodic units will be discussed in the evaluation section (6).

Due to the importance of disfluent segments in the prosodic phrasing task as explained above, annotators couldnt ignore their presence. However, prosodic phrasing task is in itself sufficiently complex and we asked annotators to focus on it; we could not have asked them for another task. In order to distinguish between boundaries of prosodic level and other boundaries associated with disfluencies (see [30], that show a very good inter-agreement for boundaries of disfluencies while they recognize that these boundaries fall within another dimension) we ran another annotation task specify to disfluencies.

3.5 Quantitative Overview

A more detailed evaluation is presented in the section 6. Here, we simply use the mean of the number of items produced by the two annotators (all the dataset was at least double-annotated). More precisely, in figure 6 we report the figures for each kind of units created. The distribution of the size of the prosodic units in terms of tokens is given in figure 7.

.

In Figure 7, `pu-disf` correspond to prosodic units hosting at least one disfluencies while `pu-r` have more of regulatory nature[5]. The later have been automatically determined from a fixed set of lexical items that are extremely frequent in these functions (*ouais, mh, ok, d'accord...*). More

[5] The term is inspired by the BDU-Regulatory of [37].

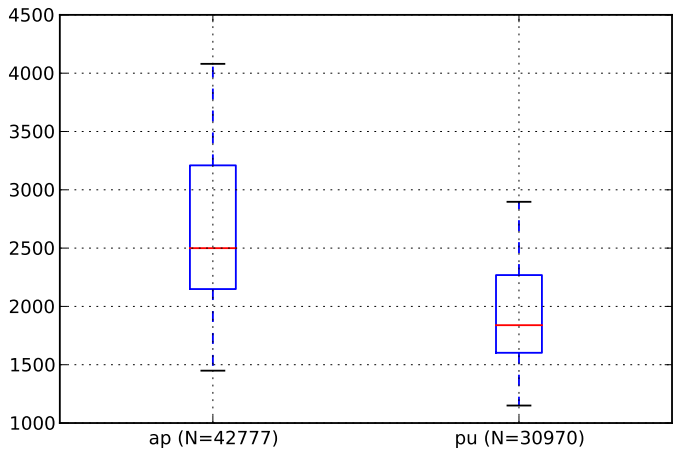

Fig. 6 Variation of figures according to speaker

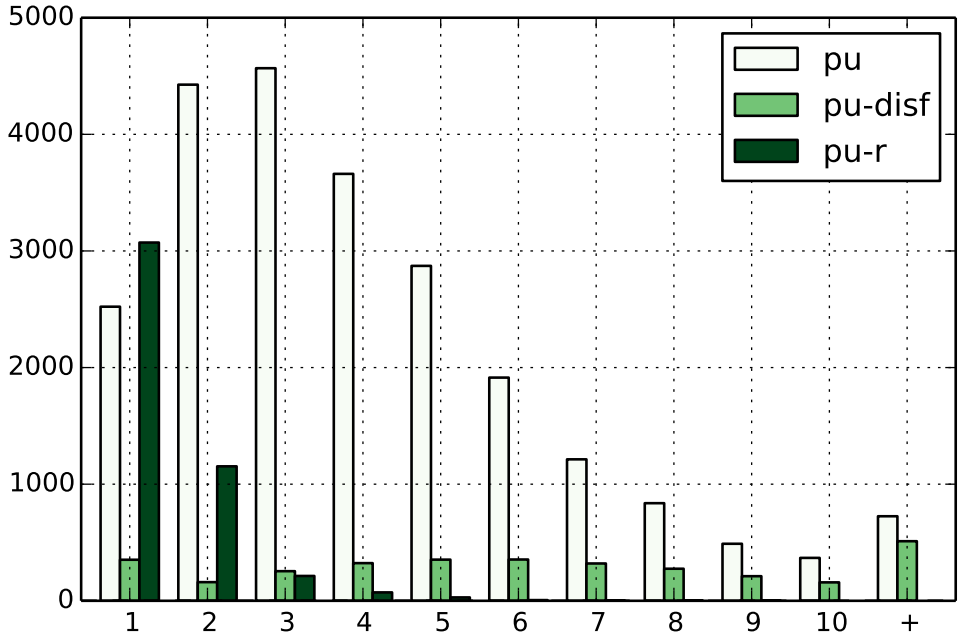

Fig. 7 Distribution of prosodic units (`PU` = `ip+IP`) lengths (in tokens)

precisely, to consider a `pu` to be a `pu-r`, the transcription of the units must feature only tokens from the list (1). The vast majority of the `PU` categorized this way correspond to feedback utterances (See for specific studies on this kind of units [88]).

(1) mh, ouais, oui, okay, d'accord, voilà, bon , @[6], non, ah, euh, ben, et, mais, *[7], heu, hum, donc, eh, beh, oh, pff, hein

As it can be seen in Figure 7, PU made of 2 to 4 tokens are the most frequent ones. The PUs composed of 1 token are in majority feedback (regulatory) items. The ones made of 2 to 4 correspond to frequent syntactic structures such as *Pronoun + (Aux) + Verb* (see Section 5 for details on syntactic dimension of the dataset) in spontaneous speech. Left-dislocation are also frequent and tend to be segmented as a major unit by the naive coders. All these structures are more frequent in our data than in other types of spoken data and tend to make our major prosodic units shorter (in number of tokens and therefore duration) than expected. We observe a drop in the distribution at length of 5 for the non-disfluent units while proportion of disfluent ones increases. This provide a first idea of what are the basic prosodic units in this kind of corpus.

4 Discourse Units Segmentation

Concerning discourse units, the annotation campaign also involved naive annotators that have segmented the whole corpus. This was realized thanks to a discourse segmentation manual, inspired by [71] but adapted to our interactional spoken data and simplified to be used by naive annotators. The manual combined semantic (eventualities identification), discourse (discourse markers) and pragmatic (recognition of specific speech acts) instructions to create the segmentation. Such a mixture of levels has been made necessary by the nature of the data featuring both rather monologic narrative sequences and highly interactional ones. Manual discourse segmentation with our guidelines has proven to be reliable with κ-scores ranging between 0.8 and 0.85 (See section 6 for details).

4.1 Discourse units definition

Discourse segmentation has been specifically addressed both for written and spoken data including monologues and dialogues, but with different focuses. Generally speaking, it concerns at least two levels of units: utterance- or clause-like units vs. paragraph of topic-like units. The latter is the subject of a vast natural language processing literature both for written data [53] and spoken data [77]. We focus here on the former that has been of interest for semanticists and discourse analysts as their basic unit analysis, sometimes called elementary discourse unit [84]. Relational approaches of discourse have used them as their basic blocks for building discourse structure, such as sentential units [55].

The discourse units we wanted to annotate correspond to the ones defined by [83] as a units that communicate information about not more than one event, event-type or state of affairs in a possible world. There are therefore the rough semantic counterparts of independent clauses in discourse. However, interactional and dialogic aspects require their definition to take into account the conversational notion of turn as it is explained in the paragraph below. There are also closely related to the macro-syntax *Illocutionary Units* [65] but approaching these units from a more semantic-pragmatic (rather than syntactic) as explained below.

As explained above we took a rather semantic view on the definition of a discourse unit. We combined semantic criterion of Vendlers style eventualities identification [104, 8], discourse criterion (presence of discourse markers) and pragmatic criterion (recognition of specific speech acts) to define our discourse units. We adopt here a very large definition of discourse markers [96] that includes discourse connectives but also adverbials in IP-adjunct position or sequence of them.

[6] The symbol for coding laughter in our dataset.

[7] The symbol '*' is used for noises; it includes voice related noises such as *clearing throat* but also other un related noises.

However, we specified in the guidelines how to use their presence. Discourse adverbials and conjunctions are proposed to be very good clues of discourse boundaries and are including in the discourse unit they introduce as illustrated in example (4). Spoken particles like *en fait, quoi* (in fact, what) are also interesting clue but of termination of the previous unit (and are therefore included in their preceding unit). Maore attitudinal markers (e.g *tu sais / you know*, *je crois / I think*) could constitute specific discourse unit but are instructed to be ignored since they can be easily automatically handled later on. Finally, markers that correlates well with reported speech introduction *ah /ah, mais / but* are also mentioned since we segment reported speech and its introducer.

We also made use of Stedes definition as a basic but solid semantic viewpoint of Elementary Discourse Unit [99, p89]: *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 events, serving as a complete, distinct unit of information that the subsequent discourse may connect to'.*

To sum-up our discourse unit is a segment describing an eventuality like example (2) or a segment bearing a clear communicative function (example (3)). This case shows to which point our units are close to Rhapsodie's *Illocutionary Units*. Indeed, it is the speech acts that we ask the annotator to segment. This is crucial in our corpus since relatively narrative sequences alternates with fiercely dialogic sequences featuring feedback, short answers and other fragments [**?**] typical from interactional speech. In the first two cases, the identification of a main verb is a strong clue. The segment generally includes all the arguments of this verb (2-b) except if strong discourse clues such as discourse markers are signaling some discourse articulation between the arguments and and the remaining of the clause (example (4)).

(2) **Eventualities**

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

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

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

(3) **Clear Communicative Function**

a. [Locuteur1: Tu vois où c'est?]$_{du}$ [Locuteur2: oui]$_{du}$
[Speaker 1: You know where it is?]$_{du}$ [Speaker 2: Yes]$_{du}$

b. [Locuteur1: Je ne voulais pas les déranger]$_{du}$ [Locuteur2: oui bien sûr]$_{du}$
[Speaker 1: I did not want to disturb them]$_{du}$; [Speaker 2: Yes of course]$_{du}$

(4) **Discourse Markers inducing segmentation**

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

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

Moreover we distinguished between several units in discourse: *discourse units* and *abandoned discourse units*. The later are units that are so incomplete that it is impossible to attribute them a discourse contribution. They are distinguished from *false starts* (that are included in the discourse unit they contributed) by the fact that the material they introduced cannot be said to be taken up in the following discourse unit.

(5) a. **Abandoned discourse unit**
[et euh mh donc t(u) avais si tu veux le sam- + le]$_{adu}$ [pour savoir qui jouait tu (v)ois]$_{du}$
[and err mm so tu had if you want the sat- + the]$_{adu}$ [in order to know who play you see]$_{du}$

b. **False start**
[c'é- je crois j'étais heu je me rappelle plus si j'étais on était au lycée ou en (...) en première année de fac je crois]$_{du}$
[it w- I think I was err I do not remember whether I was we were in high school or in (...) first year college I think]$_{du}$

4.2 Annotation process

Concerning the annotation of discourse units, it was not as complicated as in the case of prosodic units. The annotation campaign involved naive annotators that have segmented the whole corpus. This annotation was performed without listening to the signal but with timing information. Although, not listening to the signal can seem to be awkward choice, it has been made after deep discussion between the expert and project members. Since the key objective of the datasets is to study the associations between prosodic, discourse and syntactic units, our intention was to keep as independent as possible criterion from segmenting our prosodic and our discourse units. Pilot experiments within the experts group have been performed to assess the possibility of such an approach. During the segmentation, not being able to listen leads to more ambiguities in the overall interpretation of the message than there are in reality. When annotators cannot decide between two segmentations and feel that they should listen to the units, we instructed to segment according to the most standard and simple interpretation. This segmentation is therefore not meant to be more or less correct that one that would be made while listening to the signal; it is a different segmentation. The one one can produce only using text surface cues together with pauses information. This turned out to work relatively well since the inter-coder agreement is satisfying as we will discuss in the evaluation part.

The annotation was performed with Praat [21] but without including the signal window, only the time-aligned word tiers. The tiers provided were the IPUs from both speakers, the corresponding tokens and two empty tiers for performing the annotation. The Discourse Units (`DU`) boundaries were instructed to be anchored on token boundaries. As a consequence, IPU can be seen as a superfluous potential source of bias, however simply reading the tokens sequences is rather tiring and time consuming over large period of time because need of constantly adapting the zoom level to be able to read the tokens. IPUs on the other hand, with their bigger size are more convenient for reading. The segmentation was performed by adopting a set of discourse segmentation guidelines, that closely correspond to the definitions proposed in subsection 4.1 and was inspired from [26, 71] in terms of the style of instructions provided.[8]

Practically speaking, a discourse unit (DU) consists of a main predicate, and all its relating complements and adjuncts as illustrated in (2-b). Additional cues such as discourse connectives articulating discourse units were also used. Finally, mainly because of the interactive dialogic phenomena, e.g. question-answer pairs, we added a pragmatic criterion (basically allowing speech acts of any surface form to constitute a discourse unit) for allowing short utterances to be acceptable discourse units [50], *e.g. "yeah"*, and sentence fragments, *e.g. where?*. We originally allowed for discourse embedded structures (also supported by [99] definition) but they were rarely used by the annotators and yielded low inter-rater agreement. We decided therefore to work for the time being with a flat segmentation.[9]

[8] See the resource page for consulting the guidelines.

[9] This does not mean that we do not think that embedding structures are important or interesting. But our naive annotators with our guidelines did not manage to mark them consistently. Indeed, even if this task can be

Iterative creation of the guidelines The creation of the guidelines had been an iterative process. Starting from [71], a discourse annotation manual for written text, we modified the manual by removing rare cases in spoken language and adding specific spoken phenomena (such as turn alternation that plays a role in the definition of the units). We used this first version of the manual to segment 10 minutes of conversation. We then updated it and run a first annotation round with four annotators working on 15 minutes of 2 different files. A debriefing session was organized and the segmentations were checked. This session provided the annotators with much more examples they will use intuitively later. A second annotation round was performed on one hour of data. Again a long debriefing session was organized. After that, the annotators worked independently on the data. The annotation period was about 2 months for annotating a little more than 4 conversation of one hour. All the data is at least double-segmented and some parts have up to 4 concurrent annotations. The annotation took between 10 to 15 times the real time.

Annotation tool The choice of Praat as a segmentation tool was a pragmatic one. Our students know how to use it and therefore there no training time on that side. Moreover, the other candidates we considered (ANVIL [61] and ELAN [108]) did not seem to offer the possibility of using only time-aligned transcription without the signal.

Description of the tiers We planned originally two tiers, one for the base discourse units and one for handling discontinuities generated by parentheticals and disfluencies. Indeed, these phenomena are able to be inserted within a discourse without necessarily splitting it functionally. A single tier is not able to represent such structure (at least if no mechanism such as joining relation is provided like in [26]). Theoretically, two tiers are therefore necessary. However, in practice coders used rarely the possibility of discontinuous units and with very poor agreement. We therefore simply ignored this tier at the end. Such decision has some impact on the final dataset but it should not be significant given the low numbers of segmentations performed on this second tier (less than 1% of the number units proposed in the main tier).

Disfluencies Disfluencies were split into two cases: *abandoned discourse units* and other disfluencies. Only the former resulted in specific units. *Abandoned discourse units* are described above. Other disfluencies are instructed to not split the discourse units and therefore have no impact in our discourse unit segmentation. A specific disfluency annotation on which we do not develop here was carried on in parallel (See [74] for preliminary results).

4.3 Narrative sequences

In addition, a segmentation of each conversation into *narrative sequences* was performed. This work was guided by the need to separate two types of sequence in our data: (i) *narrative sequences* where one of the participant is the main speaker who is telling a story, (ii) *transition sequences* where both participants have more symmetric roles and in which they negotiate the next topic or story to develop. At the theoretical level, these narrations are both related to Labov's narration [64,63] and to complex discourse units formed of narrative sequences from more recent discourse analysis work [83,6,5]. The annotation was made intuitively simply by asking annotators to mark the boundaries of the main stories in our conversations. This was performed by one naive annotator and checked by one of the author. No inter-annotator agreement measurements was performed.

automatized for written data [2], it is much more complicated in the absence of punctuation. Crucially, embedding structure involve prosodic structuring and we did not give a access to prosodic information in the annotation process as explained above.

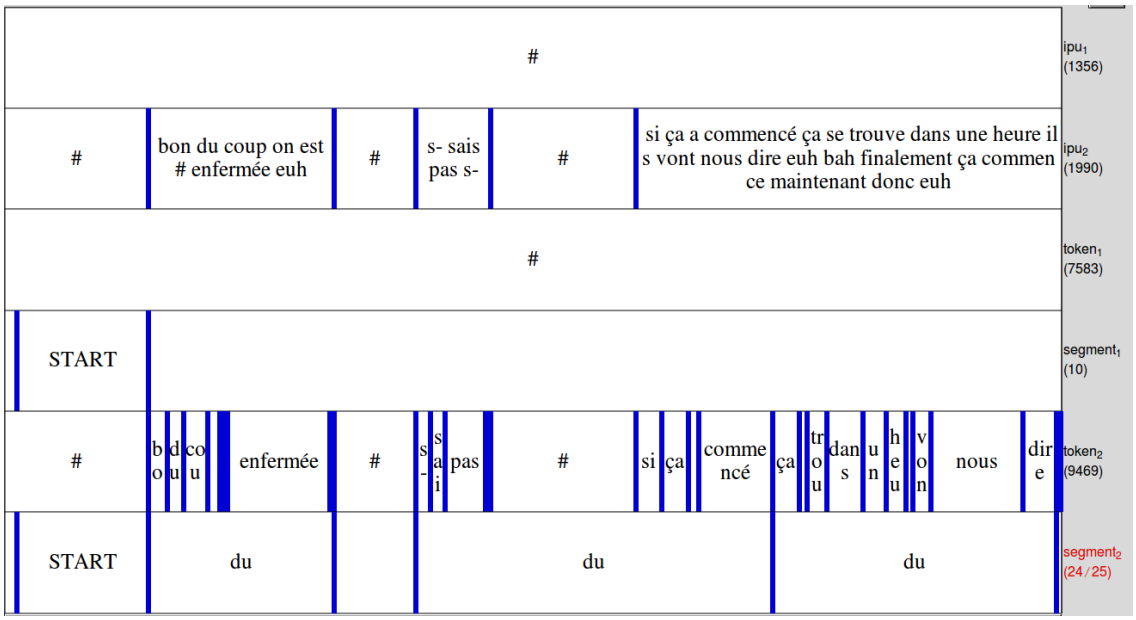

Fig. 8 Discourse Units Annotation

4.4 Quantitative Overview

A more detailed presentation and analysis of the annotation figures is presented in the section 6. Here, we use simply provide basic statistics for the items produced by the two annotators. More precisely, in table 1 we report the figures for discourse units (`DU`), abandoned discourse units (ADU) and narratives sequence (Narr). The distribution of the size of the discourse units in terms of tokens is given in figure 9.

	Abandonned Discourse Unit	Discourse Unit	Narration
Min	55	689	2
Max	317	1336	17
Mean	131.7	920.3	9.25
Std. Dev.	80.2	190.7	3.9
Total	2043	15463	148

Table 1 Basic quantitative figures for discourse annotation

In figure 9, `adu` refers to *abandoned discourse units*, `du-disc` to discourse units hosting a disfluency and `du-r` were discourse units composed only of a specific set of discourse markers as already explained in page 8.

We observe in Figure 9 that short DU are mostly related to feedback. The remaining short `DU`s can be explained by reported speech introducers (that are segmented as independent `DU`s) as well as short repetitions (that functionally can often be associated with feedback also). On the opposite, the disfluent units are over-represented among the long `DU`. This is expected since guidelines specifically mentioned that disfluencies (that do not prevent some kind of utterance completion) were included in the `DU`. Disfluencies themselves are often made of 3-4 tokens to which the rest of utterances need to be added. Finally, for the medium size (4-8), `DU`s are dominant and the distribution of the `DU`s length in this category is relatively even. This provide a first idea of what are the basic discourse unit in this kind of corpus.

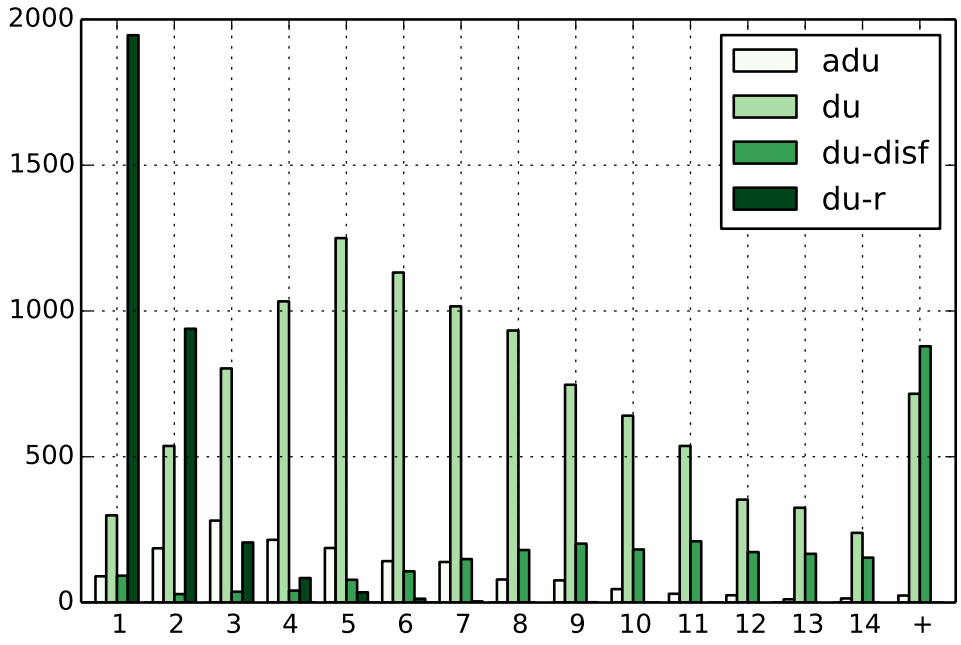

Fig. 9 Distribution of discourse units lengths in tokens (Short : 0-3; Medium : 4-9 ; Long : 10+)

5 Syntactic automatic enrichment and correction

5.1 POS-tagging

The morphosyntactic tags of the transcriptions have been obtained in three steps. In a first step, a stochastic tagger trained on written french texts was modified and applied to the transcriptions of the CID corpus. In a second step, an error analysis was performed on the output tags. Phenomena specific to spontaneous speech in interaction were identified and included in a new version of the tagger model. In a final step, a manual correction was performed on these new output tags for the 115, 000 tokens of the whole corpus. This manually corrected version was afterwards used as a gold standard to evaluate the performance of our tagger adapted for spontaneous speech transcription input.

5.1.1 Adaptation of the LPL written french tagger

The LPL tagger [90] is a stochastic tagger trained on written french texts. The tagger executes the three standard following operations. In first, a rule-based tokenizer splits the raw textual input in a sequence of tokens. In a second step, a morphosyntactic lexicon allows to associate to each orthographic token form its corresponding tag distribution. An example of tag distribution is illustrated table 2. The last operation is the desambiguisation task. It consists in selecting among the set of sequences generated by the combinatory of form ambiguity the tag sequence with the highest probability.

The probability of a sequence of tags is computed thanks to a stochastic model using the Hidden Markov Model machinery. It makes use of the conditional probability distribution of tags given a left context (e.g. the previous tags). The model thus consists of a list of *patterns* defined by the tag sequence of their left context and their associated conditional distribution of tags. An example of pattern is given table 3. The patterns of our model are extracted from the *GraceLPL*

form	lemma	sampa	tag	features	probability
bon	bon	bo∼	`Adjective`	Afpms-	0.805
bon	bon	bo∼	`Adverb`	Rgp	0.190
bon	bon	bo∼	`Noun`	Ncms–	0.005

Table 2 Tag distribution for the french lexical form *"bon"* (meaning *"well"* in its adverb use, *"good"* in its adjective use, or *"voucher"* in its noun use). Features encode the fine-grained morphosyntactic description of the tag, "Ncms" stands for example for common noun with singular number and masculine gender.

corpus, a version of the *Grace/Multitag* corpus (see [75]) which contains about 700,000 tokens with morphosyntactic annotation following the tagset features Multext [57]. *GraceLPL* is regularly corrected and enriched in order to improve its tagging.

left tag context	following tag	conditional probability
`Verb Adverb Determiner`	`Adjective`	0.09
	`Adverb`	0.01
	`Noun`	0.90

Table 3 The pattern identified by the sequence of tags `Verb Adverb Determiner` can be followed by three possible tags : `Adjective` with a probability 0.09, `Adverb` with a probability of 0.01 and `Noun` with a probability of 0.90. Other tags have null probability to occur in this context.

The morphosyntactic information has been organized in an ad-hoc way in 48 tags[10] (2 types of tags for punctuation marks, 1 for interjections, 2 for adjectives, 2 for conjunctions, 1 for determiners, 3 for nouns, 8 for auxiliaries, 4 for verbs, 5 for prepositions, 3 for adverbs and 15 for pronouns). The pattern model describing our tagger is composed of 2,841 patterns of varying size (the largest left context in the list of patterns counts 8 categories). The evaluation of the model is performed by comparing the tagged output with the reference. For the selected tagset of 48 categories mentioned above, the performance of the tagger (version 2011) reaches a score of 0.974 (F-measure). This value corresponds to a tagging error rate of 2.56%, to compare for example with an error rate of 10.7% obtained when using solely the morphosyntactic lexicon frequency information (see [20]).

In order to tag our corpus, the enriched orthographic transcriptions were filtered of annotations not containing syntactic content (filled pause, hesitation, truncation, laughter, ...). The tagger has been also modified to account for the absence of punctuation marks in the input transcription. The input was therefore segmented by isolating segments of text separated by pause duration greater than 500 milliseconds. Within each of these segments, it was moreover allowed to the tagger to insert punctuation marks when appropriate (i.e. when this insertion increases the probability of the sequence of tags treated). Two classes of punctuation marks were considered, the strong one corresponding for example to full stop marking the end of the sentence and the soft one such like comma for example. This procedure gives rise to two new units defined solely on the ground of syntactic information, *pseudo-sentences* corresponding to units delimited by two strong punctuation marks, and smaller units delimited by weak punctuation marks.

5.1.2 Error analysis and a new tagger model

In a second step, an error analysis was performed on the output tags. Two major causes of errors were identified, both related on the absence of the orthographic form or of the appropriate tag entry in the lexicon. The existing resource has thus to be adapted in that way. A convenient method for investigating the problem is to create the lexicon proper to the CID corpus and to compare it with the lexicon information issued from written texts. The CID lexicon contains 6,600 different forms with a number of occurrences spanning an interval from 1 to 3,130. For each form,

[10] The tagset selection is a fine tuned balance between the size of the training corpus, its level of morphosyntactic annotations, and the informative content of each feature with respect to the grammar, see for example the recommendations in [103].

the ratio between the spoken frequency in the CID and the written frequency in texts extracted from the *GraceLPL* corpus is computed. An extract of the CID lexicon is presented table 4. Forms which are specific to spoken corpora appear in this table with an infinite ratio (i.e. their frequency in written texts is null). We listed this forms and added them to our initial lexicon. In the CID corpus, the more frequent phenomena are word reductions (e.g. *"appart"* for *"appartement"*, *"exo"* for *"exercice"*, ...), regional version or foreign words (e.g. *"cagole"*, *"strange"*, ...) and onomatopoeia (e.g. *"beh"*, *"mh"*, ...).

form	CID occurrences	ratio spoken vs written	type
ouais	2916	26035.715	DM
mais	1429	3.483	DM
quoi	1175	56.260	DM
bon	677	18.774	DM
tu sais	635	119.266	DM
...	...	...	...
beh	59	∞	missing
appart	11	∞	missing
...	...	...	...

Table 4 For each form in the CID, the number of occurrences and the ratio of frequencies between spoken versus written use. Forms were classified by type, DM for discourse marker, missing when the form entry was not in the lexicon.

Forms with a high ratio are of potential interest. It could indicate that the speakers use these forms with different and supplementary purposes than in written productions. Among those, discourse markers emerge from the CID lexicon. Spoken discourse markers may play various functions at different levels of the linguistic analysis. The discourse marker *"ouais"* (see table 4) is for example massively produced in the CID by the hearer as a backchannel during the main speaker production. Other discourse markers (e.g. *"mais"*, *"quoi"*, *"bon"*, ...) may be used by the main speaker inside fillers with discourse planning function. In the examples mentioned above, these discourse markers do not exhibit the syntactic function they fulfill in written french. At the tagger level, we decided to associate the tag `Interjection` to this special use of discourse markers[11]. We therefore modified the lexicon by adding the tag entry `Interjection` and by redefining the probabilities of the tag distribution. An example is proposed table 5. If this last modification concerns less than 40 forms of the lexicon, it has a great importance on the overall treatment since almost 10% of the tokens of the whole corpus are impacted by these changes.

form	lemma	sampa	tag	features	probability
bon	**bon**	**bo∼**	`Interjection`	**I**	**0.95000**
bon	bon	bo∼	`Adjective`	Afpms-	0.04025
bon	bon	bo∼	`Adverb`	Rgp	0.00950
bon	bon	bo∼	`Noun`	Ncms–	0.00025

Table 5 The modified tag distribution for the french lexical form *"bon"*. An additionnal entry has been added with a tag `Interjection` corresponding to the specific spoken discourse marker use of the form. The tag probabilities have thus been redefined taking into account the ratio between spoken and written use (see table 2 for comparison).

5.1.3 Manual correction and tagger performance

In a final step, the new version of the tagger was applied to the filtered transcription inputs. From this new tagged output, a manual correction was performed on the 115,000 tokens of the whole corpus. The corpus was splitted in three parts and each part was corrected by a different annotator. It took them around 200 hours in total, including the time spent to familiarize with

[11] This choice is much more motivated by technical reasons rather than linguistic ones, we had to propose a tag already present in the existing tagset with reasonably syntactic neutral properties, which is the case for `Interjection`.

the encoding features of the tagset and with the specific interface designed for the task (see figure 10). It corresponds to an amount of 25 hours of work for 1 hour of dialogue, and in average a speed of treatment of 600 tokens per hour.

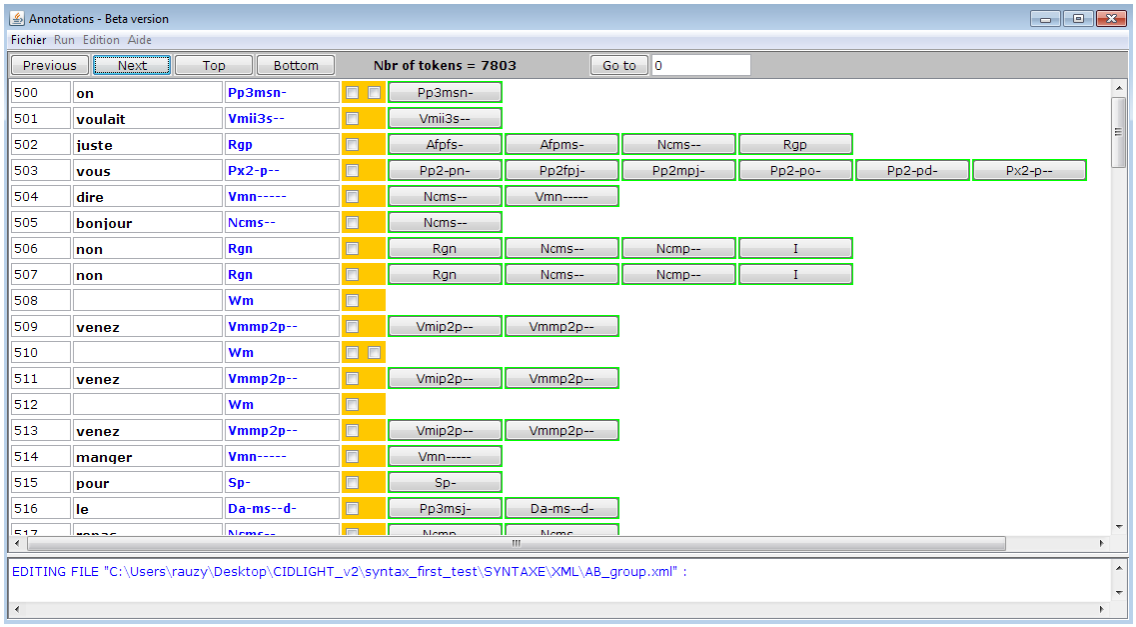

Fig. 10 The software interface used for tag correction. The transcription is presented horizontally, one line per token. Each line contains the token form, the tagger solution (second column), and the list of possible tags associated to the given form. To correct a tag, the annotator has to click on the desired tag in the list (or to type in the tag features if the entry is not proposed).

This manually corrected version was afterwards used as a gold standard to evaluate the performance of our tagger adapted for spontaneous speech transcription input. We obtained a F-measure of 0.948, or a tagging error rate of 5.2%. A more detailed inspection shows that 23% of the tagging errors are due to misclassification of discourse markers. We also investigated the location of these errors on a subsample of the corpus for which disfluency areas were annotated. It appears that 20% of the tagging errors occurs in disfluency areas. The tagger MarsaTag [92] can be downloaded at the following address : *http://sldr.org/sldr000841*. The comparison of its performance with other competing systems (for example the MElt tagger for spoken french [12] or the SEM tagger [105]), which essentially requires to adopt a common tagset and gold standard of evaluation, is out of the scope of the present paper. At this stage, the performance of MarsaTag is rather acceptable, which means that its automatic output can be already used for some studies accepting a 5% error rate for tag information. To improve its performance will require to include in the model disfluency phenomena (see for example [98]) and discourse markers classification (see for example [107]).

5.2 Chunking

Chunks can be seen as an intermediate level of syntactic processing [1]. They are the basic structures built from the part of speech tags but do not deal with long dependencies or rich constituency. Chunks are basically units centered on a syntactic head, a content word. As reminded by Abney, chunks can be related to $\phi-$sentences [48] that have a more intonational nature. An idea defended in these early works is that chunks are indeed language processing units from a cognitive viewpoint. The break-up of experimental linguistics has renewed the interest for this hypothesis and is attempting to make it more precise [16] and relate to other empirical evidences such as

eye-tracking [91]. With this idea in mind, we will investigate our prosodic and discourse units in terms of chunk size.

Analysis in chunks is an easy-to-implement and robust method. The main principle of chunking consists in including in one unit all the constituents situated to the left of each syntactic head. Chunking our data is performed by a script using 26 rules. The rules are of two types. The first type specifies tags which are always added in the same chunk as the following token as illustrated in (6). Most of the function words belong to this category.

(6) determiner (D) + anything
preposition (S) + anything
conjunction (C) + anything
personal pronoun (Pp) + anything
auxiliary verb (Va)+ anything

The second type of rules specifies an ordered pair of POS tags which must be in the same chunk (7).

(7) adjective (A) + noun (N)
demonstrative pronoun (Pd) + verb (V)
adverb (R)+ adjective (A)
proper noun (Np) + proper noun (Np)
verb (V) + verb (V)

Any number of filled pauses and truncated words can appear inside a chunk if the rules specify that the last word before these elements must be in the same chunk as the first word after them. This means that the sequence *"Determiner (D) - filled pause (FP) - filled pause (FP) - Noun (N)"* will give rise to a single chunk: DN. Otherwise, they are attached to the beginning of the next chunk. The same approach is applied to the treatment of pauses inferior to 200 milliseconds. Chunks cannot span across pauses which length is above this threshold.

We gave a category to the chunks which is in general the one of the content word present in the chunk. There are two specific cases however: (i)prepositional chunks that may include a noun or a verb ; (ii) chunk that do not have content words (because of the pause-breaking rule) that we are calling *Interactional Chunks*. This approach provided us 8 chunk types: the standard VC *(Verbal chunks)*, NC *(Nominal chunks)*, AC *(Adjectival chunks)*, RC *(Adverbial chunks)* PC *(Prepositional chunks)* as well as IC (*Interactional Chunks*), DisfError (*Disfluencies and POS errors*), PVC (*Preposition + Verb chunks*).

5.3 Quantitative Overview

In order to provide a general idea of the syntactic related annotation we provide the normalized distribution of the POS-tags and of the chunk types. We also compare them with the *GraceLPL* corpus (a 700,000-word corpus mainly composed of newspaper articles and classic French litterature) processed with the same chunking rules (See Tables 11 and 12).

The POS and chunk type distribution are strikingly different. Some important differences were expected: Interjections (I), filled pauses (FP), truncated words (WF) are much more frequent or present only in spoken data. There are also more interesting elements: (i) verbs, personal and demonstrative pronouns are much more frequent in the spoken dataset; (ii) nouns (including proper nouns), adjectives and adverbs are much better represented in the written data; (iii) determiners and prepositions are also more frequent in written data; (iv) conjunctions are more frequent in spoken data ; (v) negation adverbs are also more represented in spoken data.

The first difference (i) concerning relative high frequency of verbs and pronouns is a first hint at the nature of the production. Spontaneous speech tends to have many very short utterances

Tag	POS	Spoken	% Spok.	Written	% Writ.
V	Verb	17473	14.70	62925	9.90
N	Noun	11890	10.00	137962	21.71
I	Interjection	11186	9.41	710	0.11
Pp	Personal pronoun	10730	9.03	21057	3.31
D	Determiner	9755	8.21	97781	15.39
C	Conjunction	9064	7.63	31757	5.00
S	Preposition	7738	6.51	82957	13.05
R	Adverbe	7536	6.34	27188	4.28
Pd	Demonstrative pronoun 5891	4.96	5285	0.83	
A	Adjective	3912	3.29	47917	7.54
Po	Objective pronoun	3616	3.04	9972	1.57
Rn	Negation	3413	2.87	12000	1.89
<FP>	Filled pause	2932	2.47	0	0.00
Va	Auxiliary verb	2881	2.42	14686	2.31
Vi	Infinitive	2756	2.32	17006	2.68
Pr	Relative pronoun	1891	1.59	11206	1.76
<WF>	Word fragment	1578	1.33	0	0.00
Px	Reflexive pronoun	993	0.84	6982	1.10
Np	Proper noun	985	0.83	24299	3.82
Pt	Oblique pronoun	962	0.81	2320	0.37
P	Other pronouns	684	0.58	3112	0.49
S D	Preposition + determiner	601	0.51	17705	2.79
Pq	Interrogative pronoun	358	0.30	393	0.06
U	Unknown	39	0.03	254	0.04
Total	118864		635474		

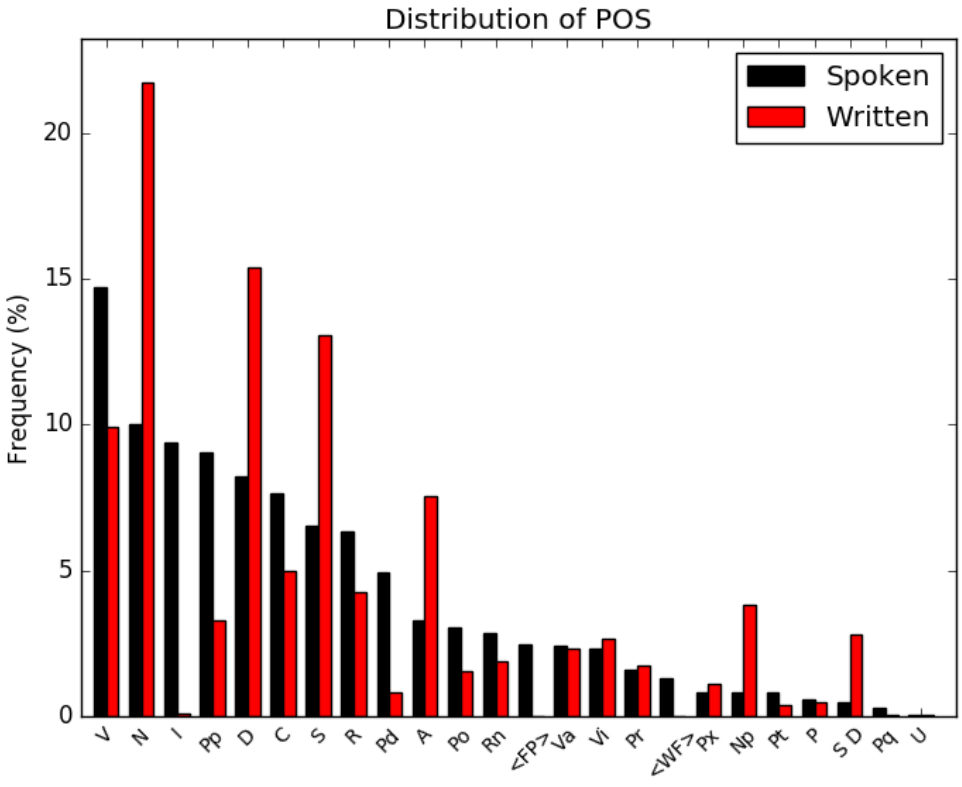

Fig. 11 Distribution of POS categories (compared to written data)

typically composed of a personal pronoun and a verb. More globally, there are much less elaborated structures such as relatives, complements and adjuncts in spontaneous spoken data than in written press genre. This also explains point (ii) concerning higher frequency of nouns, adjectives and

VC	18255	34.98	65125	21.95
NC	7704	14.76	71233	24.01
IC	6028	11.55	2206	0.74
RC	5995	11.49	19182	6.47
PC	5019	9.62	83228	28.06
DisfError	4842	9.28	10084	3.40
AC	2927	5.61	33838	11.41
PVC	1414	2.71	11737	3.96

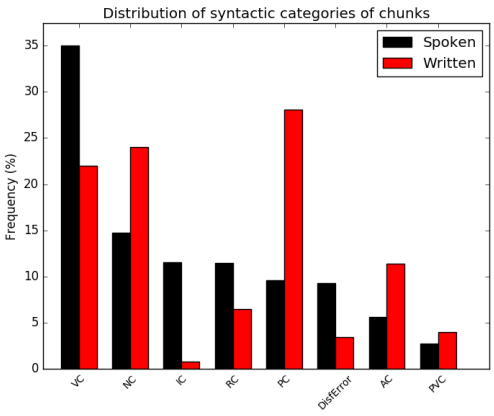

Fig. 12 Distribution of chunk categories and size (compared to written data)

adverbs in written data. They are well represented in spoken data but relatively less than in journal articles in which background information is systematically added to sentences through nominal adjuncts, numerous adjectives and adverbs. The need to always describe new information comes on top of this for proper nouns. Difference (iii) is also related to the structural complexity of the phrase, in addition to have more elements in the sentence (and therefore a higher ratio of more nouns, determiner, preposition per verb) journalistic written data overload most noun phrases with noun complements which adds again determiner and prepositions. One could be surprised by the higher frequency of conjunctions in spoken data (iv). This is partly due again to the nature of the units, conjunction mostly occur between clause-like units. Thanks to shorter simpler clauses, automatically the ratio of conjunction is higher in spoken data. Conjunctions are used to organize the relation between clauses and therefore it is expected that the inter-clausal organization of well prepared written data to be better signaled trough conjunctions. However, conjunctions (*'mais'* for example) are heavily used by speakers because of the on-line building of inter-clausal relations. Conjunctions are therefore heavily repeated and use systematically at initial positions in some sequences.

6 Evaluation

For prosodic and discourse segmentations, we calculated inter-annotator agreement through Cohen's κ [29,24,4] between experts and naive annotators as well as among naive annotators. For evaluating multiple coders' agreement we used standard multi-κ measure discussed in [4]. In our evaluation scripts we use κ and multi-κ implementations from NLTK package for Python [15]. During discourse and prosodic boundary annotation, for each token boundary, coders had to decide whether or not to segment. Therefore, the set of token boundaries corresponds to the set of decisions that were made and this is what serves as basis for the calculation of κ score.

6.1 Prosodic data evaluation

In Table 6 we report κ scores between each naive annotator and one of the experts in order to evaluate how close the strategy of each annotator was close to expert's strategy. The scores shown cover one hour for one speaker plus 9 minutes for another speaker of `AP` segmentation and two full hours of `PU` segmentation. The κ scores vary noticeably across annotators, ranging from 0.5 to 0.63 for `PU` and from 0.5 to 0.69 for `AP`. Such variations suggest that coders might use different strategies for segmentation.

Annotator	κ PU	κ AP
Annot A	0.611	0.614
Annot B	0.534	0.501
Annot C	0.498	0.563
Annot D	0.625	0.691
mean	0.567	0.592

Table 6 κ score of prosodic boundaries between naive annotators and an expert

Spk	AB	AC	AG	AP	BX	CM	EB	IM	LJ	LL	MB	MG	ML	NH	SR	YM	mean
Annot	C,D	C,D	B,C	B,C	A,D	A,B	C,D	B,C	A,D	C,D	A,B	B,C	A,D	A,B,D	A,B	A,D	
κ PU	0.613	0.535	0.674	0.402	0.710	0.460	0.592	0.406	0.725	0.657	0.784	0.535	0.616	0.738	0.705	0.666	0.614
κ AP	0.658	0.714	0.651	0.645	0.762	0.468	0.704	0.551	0.798	0.739	0.789	0.729	0.727	0.798	0.730	0.740	0.700

Table 7 κ score of prosodic boundaries between naive annotators (multi-κ for speaker NH)

	Spk AB	Spk CM	mean
κ PU	0.840	0.793	0.817
κ AP	0.775	0.745	0.760

Table 8 κ scores between expert annotators for prosodic boundaries

We report inter-annotator agreement between naive annotators in Table 7. The speaker NH is annotated by 3 annotators, so for this speaker multi-κ is reported. 30% of speakers for PU and 70% of speakers for AP are above 0.7 threshold. 60% of speakers for PU and 90% of speakers for AP are above 0.6 threshold. While some of the speakers have relatively low scores, most of the annotations seems are reliable. Low agreement for some speakers might indicate that their speech presents some characteristics which make it particularly difficult to annotate. Along with the naive versus expert κ discussed higher, it might also suggest that annotators understood the task of prosodic segmentation differently.

Additionally, as we had at our disposal annotations of prosodic boundaries made by two experts (covering two speakers in case of PU and one speaker plus 9 minutes of another speaker for AP), we report the inter-expert agreement in Table 8. This represents an evaluation of the cross-coding expert experiment that has been described in [73].

6.2 Discourse data evaluation

Annotator	Spk LL	Spk NH	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 9 κ scores for the discourse boundaries of each naive annotator versus expert on 15 minute fragments by two speakers.

The κ scores for the discourse segmentation are shown in Table 9. The expert segmentation of discourse boundaries covers two 15 minutes extracts of two speakers (30 minutes in total). The κ scores are high for all annotators. The annotator A is the only one whose agreement with expert is slightly below 0.8, the other annotators have very similar scores.

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 10 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 11 Multi-κ scores for the discourse boundaries on 15 minute fragments by 4 speakers, annotated by 3 or 4 naive annotators.

Inter-annotator agreement on discourse units between naive annotators is reported in Tables 10 and 11. The whole duration for each speaker was annotated by two naive annotators, the κ scores per speaker are shown in table 10. Table 11 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.

6.3 Cross-coded data

As can be seen above, there are various situations with regard to cross annotated data. To give an overview, we first report the number of `PU` for each situation of cross coding. In the column `Type` of the Table 13 the first digit is the number expert that annotated this unit, the second digit is the number of naive annotators who proposed it. However, 0 in the first digit does not mean that the annotation is not reliable since expert annotated only a small subset of the data. Such a stratification allows to select the dataset according to our needs: small dataset of very high quality could be limited to $'2?'$ or $'?4'$ while larger dataset with reasonable agreement would exclude only $'01'$ and $'10'$. A similar approach had been taken with discourse units even if expert annotation were much more limited as it can be seen in Figure 14.

Type	#
02	10609
01	8679
03	1049
21	940
22	864
20	583
11	414
10	317
12	139

Fig. 13 Cross-coding of prosodic data, Type: *xy* where x = # of experts, y = # of naive

Type	#
02	6244
01	4106
14	64
13	42
12	40
11	33
10	25

Fig. 14 Cross-coding of discourse data, Type: *xy* where x = # of experts, y = # of naive

6.4 Deeper segmentation evaluation

Segmentation tasks tend to yield skewed distributions, non-boundary being dominant to boundary cases. This is the point of κ-score to compensate agreement by chance. We consider nevertheless that simply providing a score is not enough for the evaluation, especially when applying the score to measure agreement on a new task. We therefore proposed in [78] a comparison of standard segmentation metrics with κ on our data subsets damaged systematically (by randomly adding, removing or moving boundaries as proposed in [67]). Below we only summarize the main results of this evaluation applied to our prosodic and discourse units.

The metrics that have been included in our comparison experiments are *Precision / Recall* as used for example in [100] ; *WindowDiff* (WD) [81] a specific measure designed for evaluating segmentations, which is an improved version of P_k metric [10]; *Boundary Edit Distance* [47,46] a new specific measure also designed for segmentation evaluation.

unit	#	size (tok)	size (chk)	size (PU)
token	109651	-	-	-
chunk	50298	2.18	-	-
pu	31412	3.49	1.60	-
du	18038	6.08	2.79	1.74

Table 12 Sum-up of the basic facts about Prosody-Syntax-Discourse CID Dataset

We only provide the figures for the most realistic damaged datasets: introducing near misses in prosodic (Figure 16) and discourse segmentations (Figure 15). We remind that actual score in our graphics does not mean that a given measure is more strict than another one. The only information the graphics are providing are: (i) how to compare the scores and how the score are evolving with regard to different structure of the datasets.

When *adding false boundaries*, the measures are more tolerant in the case of discourse units. This is due to the average length of units. As expected, precision decreases quickly while the decrease of recall is slower. Interestingly, WindowDiff and Boundary Edit Distance are inverted between `PU` and `DU` datasets. When *removing boundaries*, the measures decrease faster than in the previous case but the difference between `DU` and `PU` is maintained. Again, WD and BED are inverted between `PU` and `DU` datasets.

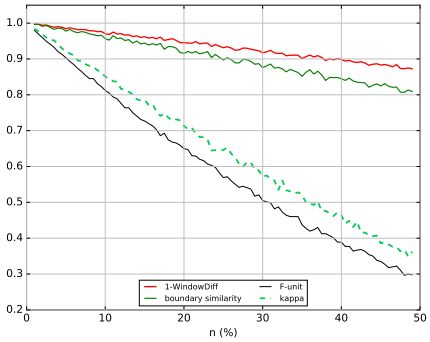

Fig. 15 Introducing near misses in discourse dataset (x=% of pertubation ; y= score)

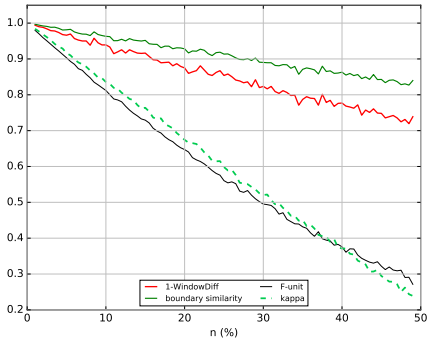

Fig. 16 Introducing near misses in prosody dataset (x=% of pertubation ; y= score)

Comparing near-miss and other errors on the `DU`, we note that structure of the data has more impact on WD and BD than the amplitude of the errors introduced. However, on a given dataset, WD and BD are efficient in capturing the differences between near-misses and other errors, BD making this difference more salient.

Finally we are comparing the different metrics across speakers on a minutes samples of the data. Figures 17 and 18 only illustrate the variation of the metrics on PU and DU across speakers. Basically, while variation is present on DU dataset it does not dramatically affect the scores. The situation is different for PU despite average values that are comparable, there is a large range of values. For some speaker, these values tend to go below reliable scores. This is in a large part due to different speaking styles. For example, highly disfluent speakers tend to yield poorer agreement scores.

7 Data overview, normalization and release

7.1 Resource overview

Quantitative overview The table 12 provided a general overview of the size of the dataset and size relationships between the different units considered. As shown in Figure 20, that shows the

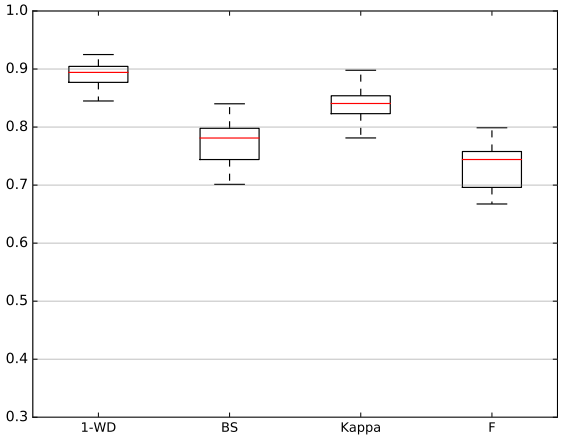

Fig. 17 Scores for DU across speakers

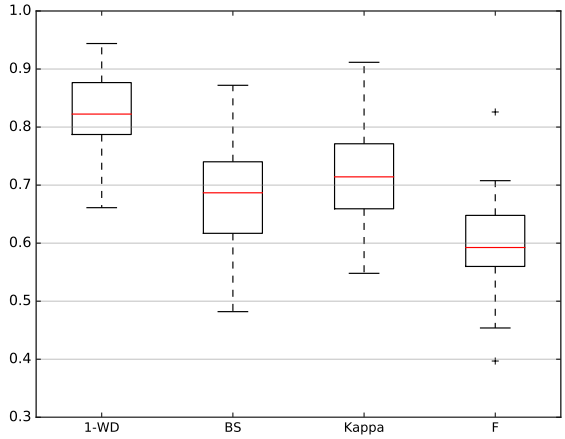

Fig. 18 Scores for PU across speakers

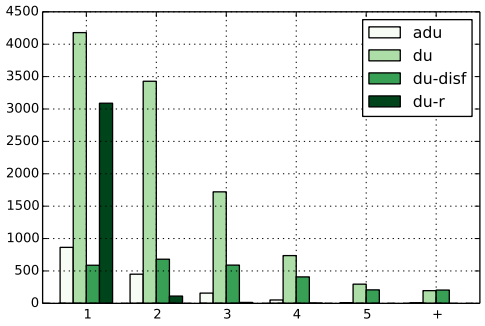

Fig. 19 Distribution of DU-sizes in terms of PU

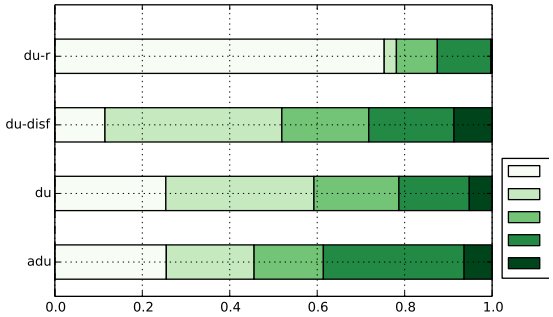

Fig. 20 Alignment between DU and PU boundaries

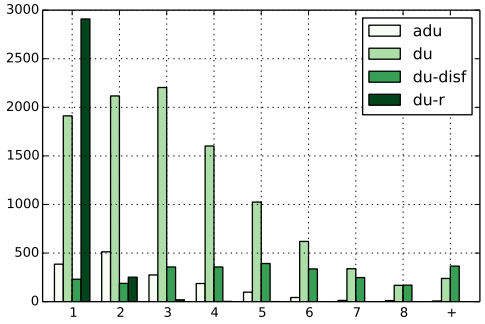

Fig. 21 Distribution of DU-sizes in terms of chunks

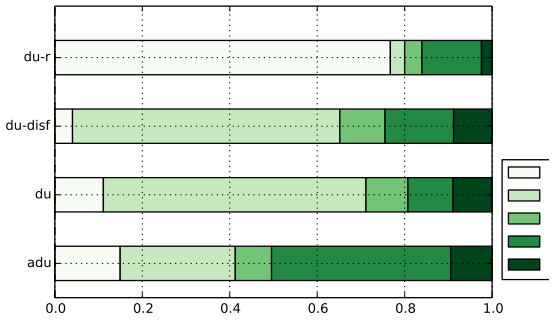

Fig. 22 Alignment between chunks and DU boundaries

alignment between left and right boundaries of PU and DU (taking DU as the starting point), prosodic and discourse units often coincide. One to one mapping account about one fourth of the data. The most frequent case is a discourse unit split into prosodic units (about one third of the cases). The other cases represented are when discourse units has some inner structure in terms of prosodic units but either left of right boundary is not matching (35%). More precisely Figure 19 illustrates the distribution of the DU-sizes in terms of PU. Figure 22 shows the alignment between

chunks and DUs while Figure 21 show the distribution of du sizes in terms of chunks. Finally, Figure 23 show the alignment between syntactic chunks and PU.

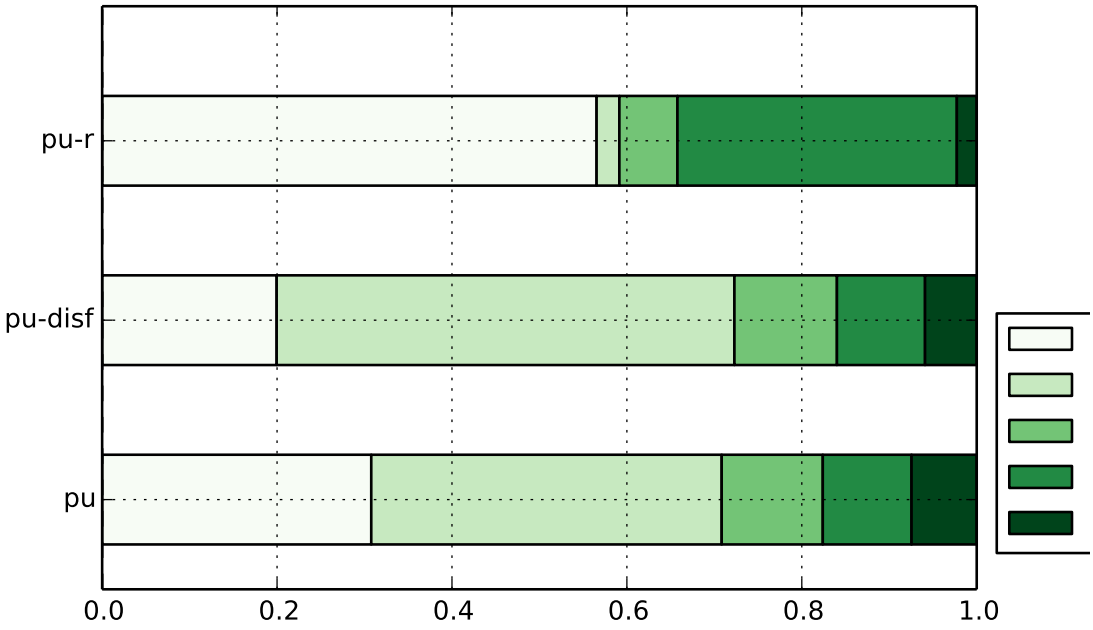

Fig. 23 Alignment between chunks and PU boundaries

Chunks are corresponding often directly to minor prosodic units. However, minor prosodic units can also split chunks into two parts. This is for example the case when the constituents have a big size such an a *Noun-Adjective* sequence in which both words include 3 syllables or more.

Qualitative illustration We present below examples of the most frequent cases introduced in the quantitative observation. In example of figure 24 we observe a perfect match between DU and PU. The most frequent case of one discourse unit split in several prosodic units is also illustrated by the longer example of figure 25. In this example, we can observe that prosodic phrasing is carving more units in the discourse segmentation structure.

Finally, examples of figures 26 and 27 illustrate mismatches. The former exhibits a rather strange IP *'dans une situation inverse c'est à dire que moi' / 'in the opposite situation that is to say that me'* in which we can clearly identify two syntactic or units. They have probably been grouped consistently by both coders on the ground of the high speech rate and in the deliberate absence of syntactic criterion for prosodic segmentation.

The later shows the difficulty of analysis some discourse markers such as the very light and frequent *'quoi'*. This example is also non-canonical with the typically final *'quoi'* moved at the beginning of the next unit. An expert coder would kept this *'quoi'* final and placed it at the end of the left unit, similarly to what is done with schwa (see Figure 2) that is holding the right boundary of the prosodic unit.

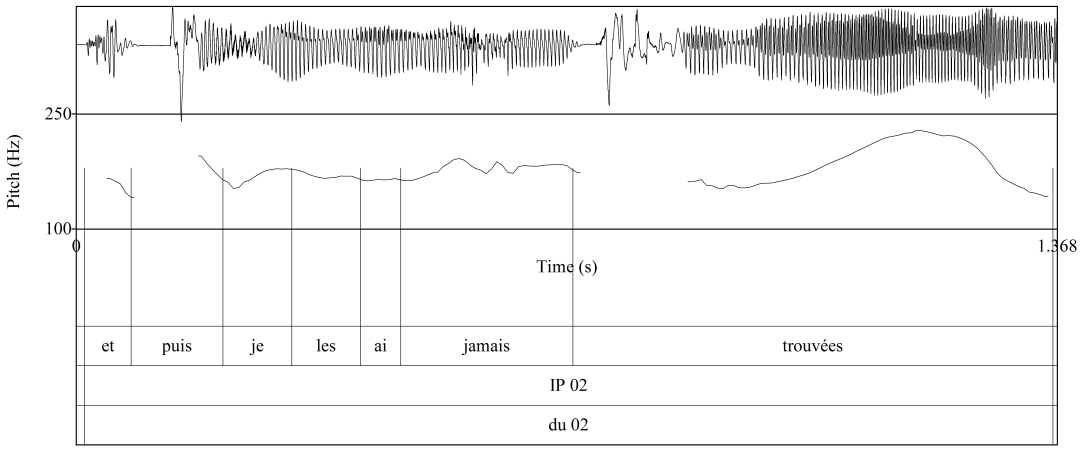

Fig. 24 One discourse unit to one prosodic unit mapping

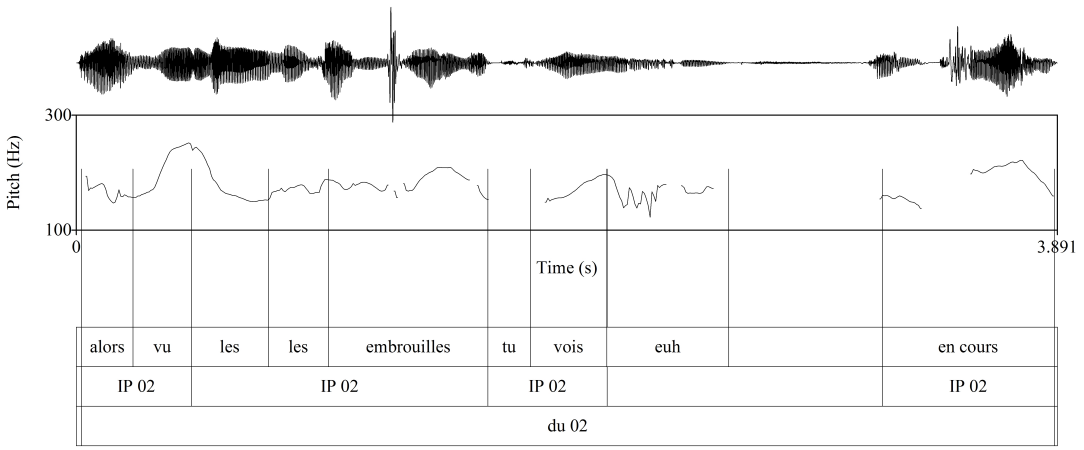

Fig. 25 Simple match between one discourse unit and several prosodic units

7.2 Discourse-Prosody Interface Discussion

Our new resource sheds a light on research fields of discourse unit and dialogue act automatic segmentation as well as descriptive studies of the prosody-discourse interface. Passoneau & Litman [77] has initiated a new trend of studies by performing both an annotation campaign, including inter-annotator agreement evaluation and setting up a system for discourse segmentation of spoken data. This system combined prosodic (mostly pause duration) and discourse connective information for approaching the proposed human reference. However, their objective is paragraph or topic segmentation instead of elementary unit segmentation. Edlund and colleagues [44] took low level prosodic information to help identifying utterance unit segmentation. This is also in line with [82] idea of the crucial importance of boundary tones for discourse segmentation. Gross and colleagues [52] used a boundary tone, a pause in speech longer than a single beat, a resetting of the pitch level, and the start of a new intonational phrase for segmenting discourse. Their attention to task-oriented dialogues led them to include the disfluencies that also introduced a break in the discourse flow. A general consensus accepted in most of the NLP literature is that low level

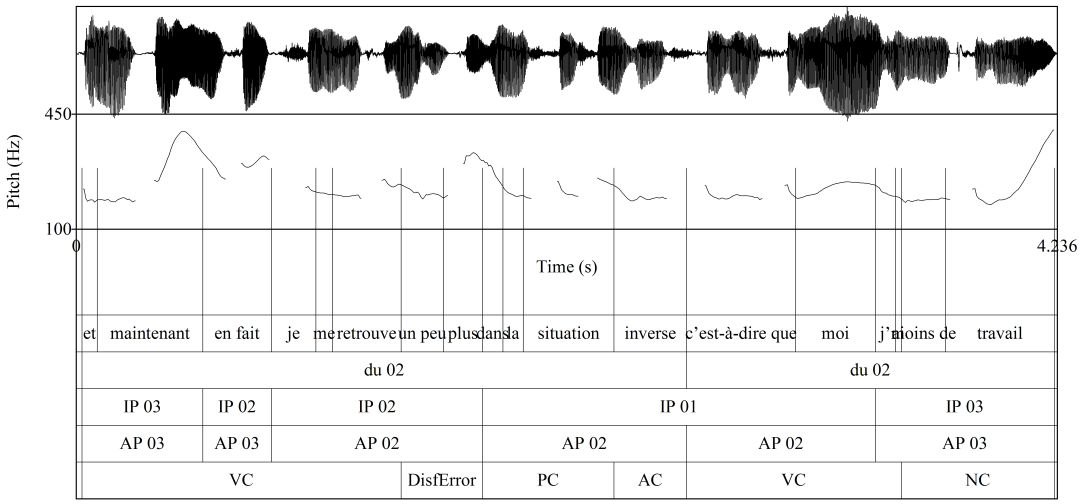

Fig. 26 Mismatch left and right

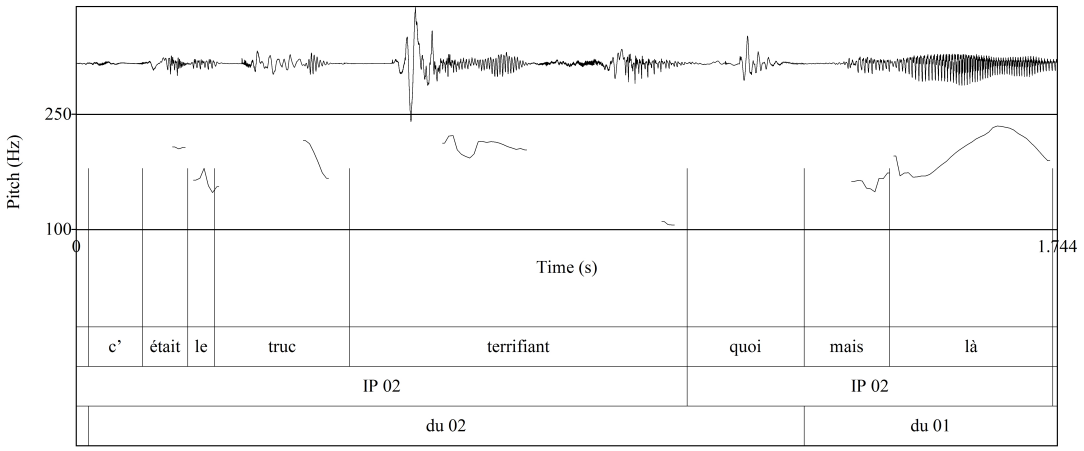

Fig. 27 Mismatch left and right (*'quoi'*)

prosodic information provides crucial information for discourse segmentation. It is often used directly as meaningful units [102]. Indeed a pilot study on our French Data [80] has also shown that even rougher prosodic analysis such as IPUs (Inter-Pausal Units for pause of 200 ms) constituted a baseline for discourse segmentation that was difficult to beat with more sophisticated methods including finer-grained prosodic analysis or including syntactic information. It is also remarkable that the language processing units of ϕ-sentences [48] have also an intonational nature. Finally, in conversational analysis, Ford and Thompson [45] have shown that intonation plays a crucial role for defining Turn-Constructional Unit [95].

More descriptive prosodists, syntacticians and discourse analysts are however much more cautious with regard to the relation of phrasing units with meaning, syntax and discourse. More precisely, we looked at how prosodic units and discourse units are distributed onto each other. In spirit, our work is closely related to those of [35,66,49] and [11]. However, our dataset has a more conversational

nature than the datasets studied in their work. About the data, [49] wanted to have an interesting spectrum of discourse genres and speaking styles while we focused on conversations both for making possible the comparative studies and to make sure to have enough coherent instances in the perspective of statistical studies. Also, while [66] requires a purely intuitive approach, we used a more balanced approach combining explicit criteria from different language domains. Finally, our annotation experiments are largely produced either by automatic tools (trained on experts data) or by naive coders. This is a major difference with the studies listed above that are based on experts annotations since it allows us scale up in data size more easily. This is opposed to Degand and Simons related work that is based on manual expert analysis and therefore concerns a relatively small dataset.

Their work is however particularly relevant for our purposes. Degand & Simon [34,35] have studied basic discourse units as a result of the combination of prosodic units and syntactic units. They observed that despite high frequency of discourse units grounded on congruent prosodic and syntactic units (called BDU-C), there can also be grouped by syntax (BDU-S) or by prosody (BDU-I). Simon & Degand [27] stated clearly in their study that following [97], neither syntactic nor prosodic completeness are sufficient for determining the boundaries of a basic discourse unit. They also show that types of BDU exhibit different distribution across discourse genre and propose that these types correspond to different discourse strategies. According to [36] and [27], BDU-C correspond to simple and neutral way to present information; BDU-I are used to create an informational macro-unit and BDU-S correspond to a more emphatic style, or resulting from careful discourse planning.

Finally, regulative BDU (or *BDU-R*) are units that contains syntactically floating elements such as phrasal adverb, connective or discourse marker but still isolated by prosodic boundaries. Our PU-R and DU-R constitute in our data a large part of those BDU-R. While [27] reported about 10% of this kind of units, even in conversational setting, in our data our *DU-R* (which is a rather low estimate for BDU-R) amounts for almost 20% of our discourse units. This underlines the fiercely interactional nature of our data despite its rather narrative nature. Our spontaneous data also tend to have many very short utterances composed typically of simple pronouns and a verb (as mentioned in section 5) which explains the relative small size of units in terms of tokens.

Beliao and colleagues [11] have a very similar approach in terms of synchronized boundaries that are both illocutionary (discourse) units (IU) boundaries and intonational period (IP) boundaries. However, the way these units are defined differs strongly from ours since the ratio IU / IP is reversed compared to our DU / PU. Their period is indeed defined to be much larger than our PU.[12] To identify a period boundary all the following conditions must be verified: pause of at least 300ms, significant F0 movement, and pitch reset. [85] argued for a phonological structuring of spontaneous speech into several phrasing levels similarly to read speech. They proposed that this structuring is only affected by disfluencies and interactional processes such as turn-taking and backchannels.

To sum up the comparison of our results with these two traditions of research we can say that we support the fact that using prosodic units as discourse units (as often proposed in NLP frameworks) is indeed a very robust and interesting approach since matching boundaries are the dominant case. However, our results on PU-DU distribution also supports that a more subtle correlation between prosodic and discourse units is possible.

7.3 Resource release and diffusion

The datasets presented here are available through the Ortolang[13] platform. The primary Data are the resources `https://hdl.handle.net/11403/sldr000720` and `https://hdl.handle.net/`

[12] This is also true, to a lesser extent, of Degand and Simon works.

[13] `https://www.ortolang.fr/`

11403/sldr000027 while the chunks, disfluencies as well as prosodic and discourse segmentation are available there: https://hdl.handle.net/11403/ortolang-000918.

8 Conclusion

This paper has introduced a new set of prosodic, syntactic and discursive annotations available on Ortolang platform. We have detailed the process for creating these annotation that combines naive and expert annotation, pre- and post-processing as well as automatic annotations. We presented and discussed the evaluation of the dataset produced. Finally, we presented qualitatively and quantitatively the units of interactional data. Their distribution in fiercely spontaneous speech was still seldom known. This work however only contributes to open avenues of development for a new linguistic discipline: *Quantitative Interactional Linguistics*. Indeed, until recently, work on interactional data remained either rather descriptive (with the major contribution of Interactional Linguistics [31]) or extremely shallow. There are attempts in providing the conceptual and technical framework for going in this direction as in the work of DiAML (Dialogue Markup Language) [22] rea for It is a real challenge for the formal and computational linguistics to approach this kind of data but it is a prerequisite to allow for systematic coding and quantitative analysis. In the paper, we have not escaped the difficulties but have shown that this is a possible move, and in fact, a necessary one if one want to address the dominant situation of language use, despite its lack of representation in language corpora available.

Acknowledgements This research has been funded by the OTIM project, ANR grant number ANR-08-BLAN-0239 and further supported by the EQUIPEX Ortolang (Grant Number: ANR11EQPX0032). We would like to thank all the OTIM team (and specifically Robert Espesser and Brigitte Bigi for hard work around the data) as well as Cristel Portes for helping on the prosodic segmentation guidelines. We also would like to thank our team of annotators: Ronan Cardinal, Marie Massot, Maud Martinez, Fanny Gavila, Héloise Schneider, Marion Pignoly, Mathilde and Adèle Desoyer, Antoine Amir, Julie Agu and Anita El Hajj.

References

1. Abney, S.: Parsing by chunks. In: R. Berwick, S. Abney, C. Tenny (eds.) Principle-Based Parsing. Kluwer Academic Publishers, Dordrecht (1991)
2. Afantenos, S., Asher, N.: Testing sdrt's right frontier. In: Proceedings of COLING 2010. Beijing (2010)
3. Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al.: The hcrc map task corpus. Language and speech **34**(4), 351–366 (1991)
4. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics **34**(4), 555–596 (2008)
5. Asher, N.: Discourse topic. Theoretical Linguitics (30), 161–201 (2004)
6. Asher, N., Lascarides, A.: Logics of conversation. Cambridge University Press (2003)
7. Astésano, C., Bard, E.G., Turk, A.: Structural influences on initial accent placement in french. Language and Speech **50**(3), 423–446 (2007)
8. Bach, E.: The algebra of events. Linguistics and Philosophy **9**, 5–16 (1986)
9. Beckman, M., Ayers, G.: Guidelines for ToBI labeling, version 3.0. Tech. rep., The Ohio State University (1997)
10. Beeferman, D., Berger, A., Lafferty, J.: Statistical models for text segmentation. Machine learning **34**(1), 177–210 (1999)
11. Beliao, J., Kahane, S., Lacheret, A.: Modéliser l'interface intonosyntaxique: ratio et synchronisation entre périodes intonatives et unités illocutoires. In: Interface Discours-Prosodie Conference 2013 (IDP-2013), p. 21 (2013)
12. Benzitoun, C., Fort, K., Sagot, B.: Tcof-pos : un corpus libre de français parlé annoté en morphosyntaxe (tcof-pos : A freely available pos-tagged corpus of spoken french) [in french]. In: Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN, pp. 99–112. ATALA/AFCP, Grenoble, France (2012)
13. Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., Rauzy, S., et al.: Le cid-corpus of interactional data-annotation et exploitation multimodale de parole conversationnelle. Traitement Automatique des Langues **49**(3), 1–30 (2008)
14. Bertrand, R., Portes, C., Sabio, F., et al.: Distribution syntaxique, discursive et interactionnelle des contours intonatifs du français dans un corpus de conversation. Travaux neuchâtelois de linguistique (47), 59–77 (2007)
15. Bird, S.: NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive presentation sessions, pp. 69–72. Association for Computational Linguistics (2006)

16. Blache, P.: Chunks et activation: un modèle de facilitation du traitement linguistique. In: Proceedings of Traitement Automatique des Langues Naturelles (2013)
17. Blache, P., Bertrand, R., Biggi, B., Bruno, E., Cela, E., Espesser, R., Ferré, G., Guardiola, M., Hirst, D., Muriasco, E., Martin, J.C., Meunier, C., Morel, M.A., Nesterenko, I., Nocera, P., Palaud, B., Prévot, L., Priego-Valverde, B., Seinturier, J., Tan, N., Tellier, M., Rauzy, S.: Multimodal annotation of conversational data. In: Proceedings of Linguistic Annotation Workshop (2010)
18. Blache, P., Bertrand, R., Ferré, G.: Creating and exploiting multimodal annotated corpora: the toma project. Multimodal corpora pp. 38–53 (2009)
19. Blache, P., Bertrand, R., Guardiola, M., Guénot, M.L., Meunier, C., Nesterenko, I., Pallaud, B., Prévot, L., Priego-Valverde, B., Rauzy, S.: The otim formal annotation model: A preliminary step before annotation scheme. In: LREC (2010)
20. Blache, P., Rauzy, S.: Influence de la qualité de l'étiquetage sur le chunking : une corrélation dépendant de la taille des chunks. In: Actes de Traitement Automatique des Langues Naturelles, pp. 290–299. Avignon, France (2008)
21. Boersma, P.: Praat, a system for doing phonetics by computer. Glot international **5**(9/10), 341–345 (2002)
22. Bunt, H., Prasad, R., Joshi, A.: First steps towards an iso standard for annotating discourse relations. In: Proceedings of the Joint ISA-7, SRSL-3, and I2MRT Workshop on Semantic Annotation and the Integration and Interoperability of Multimodal Resources and Tools, pp. 60–69 (2012)
23. Calhoun, S., Carletta, J., Brenier, J.M., Mayo, N., Jurafsky, D., Steedman, M., Beaver, D.: The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. Language resources and evaluation **44**(4), 387–419 (2010)
24. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Computational linguistics **22**(2), 249–254 (1996)
25. Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., Voormann, H.: The nite xml toolkit: flexible annotation for multimodal language data. Behavior Research Methods, Instruments, & Computers **35**(3), 353–363 (2003)
26. Carlson, L., Marcu, D.: Discourse tagging reference manual. ISI Technical Report ISI-TR-545 (2001)
27. Catherine Simon, A., Degand, L.: L'analyse en unités discursives de base: pourquoi et comment? Langue française **170**(2), 45–59 (2011)
28. Chiarcos, C., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., Stede, M.: A flexible framework for integrating annotations from different tools and tagsets. Traitement Automatique des Langues **49**(2), 271–293 (2008)
29. Cohen, J., et al.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
30. Cole, J., Shattuck-Hufnagel, S.: The phonology and phonetics of perceived prosody: What do listeners imitate? In: INTERSPEECH, pp. 969–972 (2011)
31. Couper-Kuhlen, E., Selting, M.: Introducing interactional linguistics. Studies in interactional linguistics **122** (2001)
32. Cristo, A.D.: Vers une modélisation de l'accentuation en français. deuxième partie : le modèle. Journal of French Language Studies **10**, 27–44 (2000)
33. Cristo, A.D.: Accentuation et phrasé prosodique en français. Journal of French Language Studies **21**(1) (2011)
34. Degand, L., Simon, A.C.: 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 pp. 14–15 (2005). URL `http://www.univ-tlse2.fr/erss/sem05/proceedings-final/06-Degand-Simon.pdf`
35. Degand, L., Simon, A.C.: On identifying basic discourse units in speech: theoretical and empirical issues. Discours **4** (2009). URL `http://discours.revues.org/5852`
36. Degand, L., Simon, A.C., others: Mapping prosody and syntax as discourse strategies: how basic discourse units vary across genres. Where Prosody Meets Pragmatics **8**, 79 (2009)
37. Degand, L., Simon, A.C., Tanguy, N., Van Damme, T.: Initiating a discourse unit in spoken french. Discourse Segmentation in Romance Languages **250**, 243 (2014)
38. Deulofeu, H.J.: L'approche macrosyntaxique en syntaxe: un nouveau modèle de rasoir d'occam contre les notions inutiles. Scolia **16**, 77–95 (2003)
39. Di Cristo, A.: Intonation in french. Intonation systems: A survey of twenty languages pp. 195–218 (1998)
40. Di Cristo, A.: La prosodie de la parole. Solal (2013)
41. D'Imperio, M., Cangemi, F.: Phrasing, register level downstep and partial topic constructions in neapolitan italian. Intonational phrasing in Romance and Germanic: Cross-linguistic and bilingual studies pp. 75–94 (2011)
42. D'Imperio, M., Michelas, A.: Pitch scaling and the internal structuring of the intonation phrase in french. Phonology **31**(01), 95–122 (2014)
43. Duez, D.: La fonction symbolique des pauses dans la parole de l'homme politique. Faits de langues **7**(13), 91–97 (1999)
44. Edlund, J., Heldner, M., Gustafson, J.: Utterance segmentation and turn-taking in spoken dialogue systems. Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen pp. 576–587 (2005). URL `http://202.114.89.42/resource/pdf/2510.pdf`
45. Ford, C.E., Thompson, S.A.: Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. Studies in interactional sociolinguistics **13**, 134–184 (1996)

46. Fournier, C.: Evaluating text segmentation using boundary edit distance. In: Proceedings of 51st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, USA, vol. 5 (2013). URL http://www.aclweb.org/anthology-new/P/P13/P13-1167.pdf
47. Fournier, C., Inkpen, D.: Segmentation similarity and agreement. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 152–161. Montréal, Canada (2012)
48. Gee, J.P., Grosjean, F.: Performance structures: A psycholinguistic and linguistic appraisal. Cognitive psychology **15**(4), 411–458 (1983)
49. Gerdes, K., Kahane, S., Lacheret, A., Truong, A., Pietrandrea, P.: Intonosyntactic data structures: The rhapsodie treebank of spoken french. In: Proceedings of the Linguistic Annotation Workshop @ COLING (2012)
50. Ginzburg, J., Fernandez, R., Gregory, H., Lappin, S.: SHARDS: fragment resolution in dialogue (2007)
51. Godfrey, J.J., Holliman, E.C., McDaniel, J.: Switchboard: Telephone speech corpus for research and development. In: Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on, vol. 1, pp. 517–520. IEEE (1992)
52. Gross, D., Allen, J.F., Traum, D.R.: The TRAINS-91 dialogues. TRAINS Technical Note 92-1. Dept. of Computer Science, University of Rochester, Rochester, NY (1993)
53. Hearst, M.A.: Multi-paragraph segmentation of expository text. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics, pp. 9–16. Association for Computational Linguistics (1994)
54. Hirst, D.: A praat plugin for momel and intsint with improved algorithms for modelling and coding intonation. In: Proceedings of the XVIth International Conference of Phonetic Sciences, vol. 12331236 (2007)
55. Hobbs, J.R.: Coherence and coreference. Cognitive Science **3**, 67–90 (1979)
56. Ide, N., Romary, L.: International standard for a linguistic annotation framework. Natural language engineering **10**(3-4), 211–225 (2004)
57. Ide, N., Véronis, J.: MULTEXT: Multilingual text tools and corpora. In: Proceedings of the 15th. International Conference on Computational Linguistics (Coling 94), vol. I, pp. 588–592. Kyoto, Japan (1994)
58. Jun, S.A., Fougeron, C.: The accentual phrase and the prosodic structure of French. In: Proceedings of the 13th International Congress of Phonetic Sciences, vol. 2, pp. 722–725 (1995)
59. Jun, S.A., Fougeron, C.: A phonological model of french intonation. In: Intonation, pp. 209–242. Springer (2000)
60. Jun, S.A., Fougeron, C.: Realizations of accentual phrase in french intonation. Probus **14**(147-172) (2002)
61. Kipp, M.: Anvil-a generic annotation tool for multimodal dialogue (2001)
62. Kruijff-Korbayová, I., Rieser, V., Gerstenberger, C., Schehl, J., Becker, T.: The sammie multimodal dialogue corpus meets the nite xml toolkit. In: Proceedings of the 5th Workshop on NLP and XML: Multi-Dimensional Markup in Natural Language Processing, pp. 69–72. Association for Computational Linguistics (2006)
63. Labov, W.: Narrative pre-construction. In: Narrative–State of the Art, pp. 47–56. John Benjamins (2007)
64. Labov, W., Waletzky, J.: Narrative analysis: Oral versions of personal experience. ? (1977)
65. Lacheret, A., Kahane, S., Pietrandrea, P.: Rhapsodie: a prosodic and syntactic treebank for spoken french. Studies in Corpus Linguistics (2013)
66. Lacheret, A., Obin, N., Avanzi, M.: Design and evaluation of shared prosodic annotation for spontaneous french speech: from expert knowledge to non-expert annotation. In: Proceedings of the Fourth Linguistic Annotation Workshop, pp. 265–273. Association for Computational Linguistics (2010)
67. Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., Zweigenbaum, P.: Manual corpus annotation: Giving meaning to the evaluation metrics. In: International Conference on Computational Linguistics, pp. 809–818 (2012)
68. Mertens, P.: International grouping, boundaries, and syntactic structure in french. In: ESCA Workshop on Prosody (1993)
69. Michelas, A.: Caractérisation phonétique et phonologique du syntagme intermédiaire en français: de la production à la perception. Ph.D. thesis, Université de Provence-Aix-Marseille I (2011)
70. Michelas, A., D'Imperio, M.: Uncovering the role of the intermediate phrase in the syntactic parsing of french. In: 17th International Congress of Phonetic Sciences, pp. 1374–1377 (2011)
71. Muller, P., Vergez-Couret, M., Prévot, L., Asher, N., Farah, B., Bras, M., Draoulec, A.L., Vieu, L.: Manuel d'annotation en relations de discours du projet annodis. Tech. Rep. 21, CLLE-ERS, Toulouse University (2012)
72. Nesterenko, I., Rauzy, S., Bertrand, R.: Prosody in a corpus of french spontaneous speech: perception, annotation and prosody˜ syntax interaction. In: Speech Prosody 2010-Fifth International Conference (2010)
73. Nesterenko, I., Rauzy, S., Hirst, D.J., others: On the probabilistic modelling of the form-function articulation for prosodic phenomena. Mathématiques et Sciences Humaines (Mathematics and Social Sciences) **180**(4), 113–126 (2007). URL http://hal.archives-ouvertes.fr/hal-00265189/
74. Pallaud, B., Rauzy, S., Blache, P.: Auto-interruptions et disfluences en français parlé dans quatre corpus du cid. TIPA. Travaux interdisciplinaires sur la parole et le langage (29) (2013)
75. Paroubek, P., Rajman, M.: Multitag, une ressource linguistique produit du paradigme d'évaluation. In: Actes de Traitement Automatique des Langues Naturelles, pp. 297–306. Lausanne, Suisse (2000)
76. Pasdeloup, V.: Modèle de règles rythmiques du français appliqué à la synthèse de la parole. Ph.D. thesis, Aix-Marseille 1 (1990)
77. Passonneau, R.J., Litman, D.J.: Discourse segmentation by human and automated means. Computational Linguistics **23**(1), 103–139 (1997)

78. Peshkov, K., Prévot, L.: Segmentation evaluation metrics, a comparison grounded on prosodic and discourse units. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (2014). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/931_Paper.pdf
79. Peshkov, K., Prévot, L., Bertrand, R.: Evaluation of automatic prosodic segmentations. In: Interface Conference 2013 (IDP-2013), p. 95 (2013)
80. Peshkov, K., Prévot, L., Bertrand, R., Rauzy, S., Blache, P.: Quantitative experiments on prosodic and discourse units in the corpus of interactional data. In: Proceedings of SemDial 2012: The 16th Workshop on the Semantics and Pragmatics of Dialogue (2012)
81. Pevzner, L., Hearst, M.A.: A critique and improvement of an evaluation metric for text segmentation. Computational Linguistics **28**(1), 19–36 (2002)
82. Pierrehumbert, J.: The meaning of intonational contours in the interpretation of discourse. Intentions in communication p. 271 (1990)
83. Polanyi, L., Culy, C., Van Den Berg, M., Thione, G.L., Ahn, D.: A rule based approach to discourse parsing. In: Proceedings of SIGDIAL, vol. 4 (2004)
84. Polanyi, L., Scha, P.: A syntactic approach to discourse semantics. In: COLING'84. Stanford, California (1984)
85. Portes, C., Bertrand, R., et al.: Permanence et variation des unités prosodiques dans le discours et l'interaction. Journal of French Language Studies **21**(1) (2011)
86. Portes, C., Rami, E., Auran, C., Cristo, A.D.: Prosody and discourse: a multi-linear analysis. In: Speech Prosody 2002, International Conference (2002)
87. Post, B.: Tonal and phrasal structures in French intonation. Ph.D. thesis, University of Nijmegen (2000)
88. Prévot, L., Bigi, B., Bertrand, R.: A quantitative view of feedback lexical markers in conversational french. In: 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 1–4 (2013)
89. Prévot, L., Tseng, S.C., Chen, C.H.A., Peshkov, K.: A quantitative comparative study of prosodic and discourse units, the case of french and taiwan mandarin. In: PACLIC (2013)
90. Rauzy, S., Blache, P.: Un point sur les outils du lpl pour l'analyse syntaxique du français. In: Actes du workshop ATALA 'Quels analyseurs syntaxiques pour le français ?', pp. 1–6. Paris, France (2009)
91. Rauzy, S., Blache, P.: Robustness and processing difficulty models. a pilot study for eye-tracking data on the french treebank. In: 24th International Conference on Computational Linguistics (2012)
92. Rauzy, S., Montcheuil, F., Blache, P.: Marsatag, a tagger for french written texts ansd speech transcriptions. In: Procceedings of the Second Asia Pacific Corpus Linguistics Conference. Hong Kong, China (2014)
93. Raymond, C., Riccardi, G., Rodrigez, K., Wisniewska, J.: The luna corpus: an annotation scheme for a multi-domain multi-lingual dialogue corpus. Proceedings of Decalog2007 (2007)
94. Rodriguez, K.J., Dipper, S., Götze, M., Poesio, M., Riccardi, G., Raymond, C., Wisniewska, J.: Standoff coordination for multi-tool annotation in a dialogue corpus. In: Proceedings of the Linguistic Annotation Workshop, pp. 148–155. Association for Computational Linguistics (2007)
95. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language pp. 696–735 (1974). URL http://www.jstor.org/stable/412243
96. Schiffrin, D.: Discourse Markers. Cambridge University Press. (1987)
97. Selting, M.: The construction of units in conversational talk. Language in Society **29**(04), 477–517 (2000)
98. Shriberg, E., Stolcke, A.: How far do speakers back up in repairs? a quantitatve model. In: The 5th International Conference on Spoken Language Processing, Sydney, Australia, 30th November - 4th December (1998)
99. Stede, M.: Discourse Processing. Morgan & Claypool publishers (2011)
100. Tjong, E., Sang, K., Déjean, H.: Introduction to the CoNLL-2001 shared task: clause identification. In: Proceedings of the 2001 workshop on Computational Natural Language Learning-Volume 7, p. 8 (2001)
101. Torreira, F., Adda-Decker, M., Ernestus, M.: The nijmegen corpus of casual french. Speech Communication **52**(3), 201–212 (2010)
102. Traum, D.R., Heeman, P.A.: Utterance units in spoken dialogue. In: Dialogue processing in spoken language systems, pp. 125–140. Springer (1997)
103. Van Eynde, F., Zavrel, J., Daelemans, W.: Part of speech tagging and lemmatisation for the Spoken Dutch Corpus, pp. 1427–1434. ELRA, Athens (2000)
104. Vendler, Z.: Verbs and times. Philosophical Review **46**, 143–160 (1957)
105. Wang, I., Kahane, S., Tellier, I.: Macrosyntactic segmenters of a french spoken corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (2014)
106. Welby, P.: Effects of pitch accent position, type, and status on focus projection. Language and Speech **46**(1), 53–81 (2003)
107. Westpfahl, S.: Stts 2.0? improving the tagset for the part-of-speech-tagging of german spoken data. In: L. Levin, M. Stede (eds.) Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop, pp. 1 – 10. Association for Computational Linguistics and Dublin City University, Dublin (2014)
108. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: Elan: a professional framework for multimodality research. In: Proceedings of Language Resources and Evaluation Conference (LREC). Citeseer (2006)