Depends on what the French say: Spoken corpus annotation with and beyond syntactic function

José Deulofeu, Lucie Duffort, Kim Gerdes, Sylvain Kahane, Paola Pietrandrea

Depends on What the French Say

Spoken Corpus Annotation With and Beyond Syntactic Functions

José Deulofeu
DELIC, Université de Provence
Aix, France.
jose.deulofeu@up.univ-mrs.fr

Lucie Duffort, Kim Gerdes
LPP, Sorbonne Nouvelle
Paris, France
luciedduffort@hotmail.com
kim@gerdes.fr

Sylvain Kahane
Modyco, Université Paris Ouest
Nanterre, France
sylvain@kahane.fr

Paola Pietrandrea
Université Roma TRE / Lattice, ENS
Rome, Italy and Paris, France
pietrand@uniroma3.it

Abstract

We present a syntactic annotation scheme for spoken French that is currently used in the *Rhapsodie* project. This annotation is dependency-based and includes coordination and disfluency as analogously encoded types of paradigmatic phenomena. Furthermore, we attempt a thorough definition of the discourse units required by the systematic annotation of other phenomena beyond usual sentence boundaries, which are typical for spoken language. This includes so called "macrosyntactic" phenomena such as dislocation, parataxis, insertions, grafts, and epexegesis.

1 Introduction

This communication presents the syntactic annotation scheme currently being developed for *Rhapsodie* a project funded by the French National Research Agency (ANR) which aims to study the syntax-prosody interface in spoken French. *Rhapsodie* aims to elaborate a freely distributed corpus, classified into different discourse genres, and doted with prosodic and syntactic annotations elaborated for the study of the relationship of prosody, syntax, and information structure in discourse.

Contrary to what is available in the anglo-saxon world, there is no freely distributed and syntactically annotated corpus of spoken French today. This is what our project aims to provide.

The only tree-bank for French, that we know of, is the Paris 7 Corpus (Abeillé et al. 2003). This is a corpus of newspaper texts, annotated mainly in Penn Tree Bank style and partially with dependency annotations, which is distributed only under highly restrictive conditions.

Some annotated corpora of spoken French nevertheless exist: The CID (Corpus of Interactional Data) (Bertrand et al. 2009) uses an annotation with typed chunks, and the VALIBEL corpus (Dister et al. 2008 ; Degand et Simon 2009) consists of delimiting maximal syntactic units. This notion, allowing segmentation of the text, is essential for any syntactic annotation, a concept we will come back to in section 2. Neither of these corpora is distributed freely and none comes close to the precision and variety of spoken language corpora existing for other languages like English or Dutch.

There is, however, an important tradition of description of the spoken French language, notably at the University of Provence in Aix, where a team led by Claire Blanche-Benveniste coined the two level distinction of "micro-syntax" and "macro-syntax" and proposed a parallel analysis of paradigmatic phenomena ranging from coordinations to disfluencies (Blanche-Benveniste 1990, Berrendonner 1990, Bilger et al. 1997, Guénot 2006, Gerdes & Kahane 2009).

Rhapsodie's innovation stems from a formalization and generalization of this tradition. The parallel annotation of prosody and syntax naturally leads to a syntactic analysis of the text as a whole, including hesitations and disfluencies, whereas other approaches tend to erase these phenomena in order to obtain standard sentences similar to written language where syntactic annotation is well-established. Examples of this latter approach include main reference corpora, for example the English Switchboard corpus (http://groups.inf.ed.ac.uk/switchboard), or the CGN (Dutch Spoken Corpus, http://lands.let.ru.nl/cgn). These types of annotation also commonly exclude phenomena such as colon effects, grafts, and associated illocutionary

units, because of their limited conception of sentence boundaries and a focus on written phenomena. The *Rhapsodie* syntactic analysis scheme tends to include all words of the corpus and finds it necessary to take account of all the above phenomena because they are, we believe, intrinsically syntactic.

The original English examples in this paper stem from the Micase corpus (Simpson-Vlach & Leicher 2006), in particular from the segment Honors Advising (http://quod.lib.umich.edu/m/micase) and from interviews that we collected ourselves (ELUI; English Language Use Interviews, Duffort in preparation). Some phenomena are specific to French and we use original examples from the *Rhapsodie* corpus; some other examples, designated as "constructed examples", are simplified constructions of phenomena we only encountered in more complex combinations.

2 Annotation

In the analysis of written text, the units of annotation are usually taken to be "graphical" sentences, i.e. the words between two periods, a neither explicit nor homogenous notion that has little or no linguistic relevance. Spoken corpus annotation, on the contrary, has to simultaneously define dependency units and the dependency annotation that we impose on these units. These two questions are not independent: The more phenomena we include in the syntactic analysis, the longer the units will become.

Our first choice concerns syntactic annotation: Functional dependency annotation has proven to be a more challenging task than phrase structure annotation but seems to be more versatile for various languages and more promising as an intermediate syntactic structure between the ordered words and semantics. All dependency based corpora have to choose a set of functions to be used in annotation. This choice is often guided by practical considerations (existing phrase structure annotation, parsers, semantic needs, etc.) but even though few have tried to give a formal and general definition of syntactic functions (Mel'cuk & Pertsov 1987), each choice of a set of functions presumes that two elements (subtrees) that share the same function have something in common: Usually this is thought to be

- the exchangeability of the two elements (at a certain degree of abstraction, excluding, for example, agreement features)
- the coordinability of the two elements

For example, to decide whether *gone* and *the bike* have the same function in *he has gone* and *he has a bike*, it is not sufficient that the two elements can be interchanged; we also need coordinability which in this case is ungrammatical. We will therefore stipulate the existence of two different functions.[1]

(1) *He has gone and a bike.

In other words, a coordination is an orthogonal construction to a head-daughter relation. This also shows in the difficulty in dependency as well as phrase structure approaches to account for coordination. The near-symmetry of coordinations violates basic assumptions of X-bar theory and head-daughter relationships. Contrary to other dependency analyses like the Prague Dependency Treebank (ufal.mff.cuni.cz/pdt) or the Alpino Dependency Treebank (www.let.rug.nl/vannoord/trees), our approach does not include coordinations in our syntactic functions, but these, as well as other paradigmatic phenomena, are encoded in what we call "piles"[2] (see section 2.3).

2.1 Dependency Units and Illocutionary Units

We don't consider that syntax can be reduced to dependency, and we have to define the delimitation of functional relations as well as the delimitation of so called "macro-syntactic" phenomena such as dislocation and colon effect that go beyond dependency. Our complete annotation therefore includes units joined by dependency, paradigmatic sub-units, and higher-level relations that are still syntactic and not purely discursive. We propose a well defined distinction between syntax based segmentation, called "dependency units" (DU), and pragmatically based segmentation, called "illocutionary units" (IU).

Applying a bottom-up approach, we first look for rectional (head-daughter) relations, which gives us the DUs: Each DU is a unit, constructed around a syntactic head that itself has no governor. We define a rectional relation using the

[1] Note that this choice is less clear in many cases, such as for example for the distinction between passives and predicative functions, or between full and light verbs.

[2] Of course this can be represented formally equivalently as a specific type of dependency, but we believe that the distinction is linguistically important and limiting the notion of dependency to true head-daughter relations makes the notion of dependency more consistent.

common criteria: i.e. constraints in terms of category, morphological features, and restructuration possibilities (commutation with a pronoun, diatheses, clefting).

In addition to these syntactic units, we define the IUs as unities that demonstrate a discursive autonomy, in other words, that have their own illocutionary force. These terms may seem surprising in formal syntax, but we believe that they are unavoidable for our task. This definition assents to traditional grammarians' intuition of sentences holding a "complete meaning" and Creissels' definition of "sentence" (2004) as a propositional content realizing an enunciation.

Both units, DUs and IUs are relatively independent and complementary and they have their own well-formedness conditions. In general, an IU is a combination of several DUs, but we will show examples ranging from simple interjections to complex embedded DUs. In some cases a rectional relation, and thus a DU can go beyond the limits of an IU.

This opposition of DU and IU reflects Blanche-Benveniste's opposition between microsyntax and macrosyntax (1990): A DU is the maximal microsyntactic unit; an IU constitutes the maximal unit of macrosyntax.

2.2 Microsyntax and Dependency Units

In this paper, we will not elaborate further on the dependency annotation itself. We have followed approaches taken by numerous other corpora such as the Prague Dependency Treebank or the Alpino Dependency Treebank (www.let.rug.nl/vannoord/trees/).

Let us consider the following utterance, typical for spoken French:

(2) moi ma mère le salon c'est de la moquette
me my mother the living room it's carpet
'My mother's living room is carpeted'

In (2), three elements—*moi,* literally 'me', *ma mère* 'my mother', *le salon* 'the living room'—are paratactically juxtaposed to a predicative unit, *c'est de la moquette* 'it is carpet'. These elements are not syntactically dependent on any element in the predicative unit. We treat them as separate DUs. We will illustrate in 2.4 the treatment we propose for the relation holding between these DUs.

2.3 Piles

Beside dependency, we acknowledge the existence of a separate mechanism of syntactic cohesion within DUs: Following Gerdes & Kahane (2009), we call the syntactic relation between units occupying the same structural position within a DU, or, in other words, holding the same position in a dependency tree, a "pile". Coordination is a typical case of piling:

(3) our two languages are {English | ^and French} (ELUI)

We consider that we also have a pile of elements occupying the same structural position in reformulations (4), disfluencies (5) or corrections (6):

(4) did a humanoid species { spring up | or exist } in various places {in the world | {not just in Africa | ^but also in Asia | ^and maybe also in southern Europe }} // (Micase)
(5) { I~ | in~ | including } kind of a general idea of these "uh" (ELUI)
(6) {I | I} have lots of other interests {like "um" | that are a little bit more like} {paleontology | ^or astronomy | ^or international religion | ^or "uh" not religion | international relations | ^so those things {I wanna & | I think I'm gonna concentrate more on} // (Micase)

Our desire to treat coordinations, reformulations and disfluencies as phenomena showing syntactic similarity resides in the fact that, as shown by Blanche-Benveniste (1990) among others, it is not always easy to distinguish between disfluency, reformulation and coordination: As an example, consider (7a), more or less interpreted in the same way as examples (7b,c) which are, respectively, a reformulation and a coordination:

(7) a. she is { a linguist | maybe a technician }
b. she is { a linguist | "um" a technician }
c. she is { a linguist | ^or a technician }
(constructed example)

In all cases of piles, we use the same notation: the segments that occupy the same syntactic position are put between curly brackets { } and they are separated by vertical pipes |. Pipes therefore separate what we call pile layers. These layers may be introduced by pile markers, usually a conjunction. If a pile marker does not play a syntactic role, it is preceded by a caret ^.

Dependencies and piles allow for a complete description of the syntactic cohesion of a DU. In (7), for example, the first layer realizes the position of attribute within the dependency structure. The syntagmatic relation between the two layers entails a paradigmatic relation between linguist and computational scientist. The second layer inherits the structural (attribute) position from

the paradigmatic relation within the dependency structure. It should also be noticed that, with the exception of abandoned layers (noted &), layers can be seen as alternatives. It is possible to walk these structures by choosing one layer of each pile, extracting as many utterances as there are paths. Each of these utterances has a complete dependency structure merely containing government and modification relations, for example, (7a) can be reduced to the two DUs in (8), which will constitute the input for the parsing process:

(8) a. she is a linguist
b. she is maybe a computational linguist

Note that *maybe*, though it acts as a pile marker, also plays a syntactic role in the context of the pile, contrarily to a conjunction (*she is or a computational scientist), the latter being marked with the caret to make this distinction.

2.4 Macrosyntax and Illocutionary Units

An Illocutionary Unit (IU) is any portion of discourse encoding a unique illocutionary act: assertions, questions, and commands (see Benveniste 1966, Searle 1976). An IU expresses a speech act that can be made explicit by introducing an implicit performative act such as "I say", "I ask", "I order". A test for detecting the Illocutionary Units that make up a discourse consists of the introduction of such performative segments (see below). A segmentation in IUs is particularly important for the study of the connection of prosody and syntax, which is the goal of *Rhapsodie*, because these units are prosodically marked (Blanche-Benveniste 1997, Cresti 2000). We use the symbol // to segment the text in IUs (but see also the symbols //+ in section 3).

It should be noted that there exist IUs that are not made up of Verbal Dependency Units. See examples (9a,b):

(9) a. SPK1: we've heard all of the "you know" big "uh" meteors coming from outer space // SPK2: right // (Micase)
b. ^and then < boom //
(constructed example)

We extend the notion of IU to a unit whose status in terms of illocutionary acts, let alone in terms of propositional structures, may be unclear, but which can form a "complete message": interjections, phatics, feed back particles like *voilà* 'that's it', *quoi* 'what', *hélas!*, 'alas', *tant pis!* 'oh well'. See for instance in the famous critical punt against French writer Corneille (10a) that could be annotated as in (10b).

(10) a. Après l'Agésilas, hélas ! Après l'Attila, holà ! (Nicolas Boileau 1828)
'After Agésilas, alas! After Attila, no more!'
b. après l'Agésilas < hélas // après l'Attila < holà //

In a context such as (11), a single IU is made up of two verbal DUs: I got up in the morning and I was with clients.

(11) I got up in the morning < I was with clients // I ate at noon < I was with clients // I went to bed at night < I was with clients //
(translation, *Rhapsodie*)

The relation between the two verbal DUs in (11) cannot be described in terms of microsyntactic dependency. Indeed, *I got up in the morning* is not dependent on the verbal construction of the following DU. Nevertheless, the existence of a macrosyntactic relation can be acknowledged. The first DU in (11), *I got up in the morning*, is not as autonomous from an illocutionary point of view: it cannot constitute a self standing message. In (11) it is not asserted that "I got up in the morning". And (11) can be paraphrased by (12a) but not by (12b):

(12) a. it is said that I got up in the morning I was with the clients
b. # it is said that I got up in the morning and that I was with the clients.

The illocutionary force of (11) is encoded by the DU *I was with clients*, which can be interpreted as an assertion even if uttered in isolation. Whereas the unit *I got up in the morning* does not have in this context any illocutionary interpretation. The subsegment of an IU supporting the illocutionary force of the IU is called the *nucleus*. It can be autonomized. The nucleus and the others segments forming the IU are called the Illocutionary Components (ICs). The ICs are always microsyntactic units and are generally DUs. The nucleus is the unit that is affected by a negation or an interrogation having scope on the IU. See for example the tests in (13) and (14):

(13) A: I got up in the morning I was with clients
B: this is not true (≈ It is not true that you were with clients, # It is not true that you got up in the morning)

(14) A: I got up in the morning I was with clients
B: Is that true? (≈ Is that true that you were with clients)
(# Is that true that you got up in the morning)

ICs preceding and following the nucleus are called pre-nuclear units (Pre-N) and post-nuclear units (Post-N). We use the symbol < to mark the Pre-N and the > to mark the post-N. These tags can be considered as explicit counterparts of commas in writing.

(15) il y a plein de trucs < tu les vois après > en fait > les défauts (*Rhapsodie*)
there are plenty of things < you see them later > actually > the faults

It is possible that, due to a particular communicative structure, the illocutionary force is carried only by a part of a DU and that the nucleus forms a DU with another IC:

(16) to my mother <+ I don't speak anymore (constructed example)
(17) two euros >+ it costs (translation, Blanche-Benveniste 1990)

The addition of the symbol + indicates that the IC on one and the other side are parts of the same UR.

3 More cases of irregularity in the interface between microsyntactic and macrosyntactic units

We will now present a number of structures that were particularly problematic for the syntactic annotation of the *Rhapsodie* corpus and that illustrate the mismatch between DU and IU boundaries well.

3.1 DU beyond the IU

Up to now, we have seen a few examples of segmentation of an IU into DUs. We will now show that there are cases, traditionally named epexegesis, where we can consider that it is in fact the DU which is segmented into multiple IU. Let us consider these two examples:

(18) SPK1: he has arrived
SPK2: last night (constructed example)
(19) She speaks French. And very well! (constructed example)

In these two examples, there are two illocutionary acts: in (18) this is evident as there are two speakers uttering two different assertions. In (19), there are two assertions. In both cases, the second illocutionary act is not (micro) syntactically autonomous. The second IU directly follows the first IU and integrates and completes its syntactic structure, being in a dependency relation with the head of the first IU (the verb *arrived* in (18), the verb *speaks* in (19)). We can therefore paraphrase the preceding examples thusly:

(20) SPK1: he has arrived
SPK2: he has arrived last night
(21) She speaks French and (what is more) she speaks French very well.

Rather than postulating an ellipsis in the second segment (as suggested by Culicover & Jackendoff 2005, among others) we analyze the two IUs as belonging to the same DU. This choice naturally descends from the modular approach we adopted, which distinguishes between illocutionary and syntactic relations. As in the case of a dependency relation crossing the IC border, we add a + symbol to indicate that the illocutionary frontier is not a limit to the DU.

In addition to dependency, piling can also cross IU frontiers, as in (22):

(22) SPK1: How often do you go {there |} //+
SPK2: {| to the States} // (ELUI)

In (22) the argument position of the verb *go* is realized twice: through the segment *there* uttered by the first speaker and through the segment *to the states* uttered as a separate IU by the second speaker. We use the notation {X|}…{|Y} when the pile between X and Y is interrupted by a syntactic frontier, in this case an IU frontier, or a discontinuity.

It should be noted that the piling in (22) does not only cross an IU frontier but it crosses a speech turn frontier as well, as it is realized by two different speakers. We do not consider the speech turn as a limit for the extension of syntactic phenomena, rather we assume that there can be co-construction of semantic content and syntactic structures in dialogues

3.2 Inserted IUs

An IU can be inserted into another IU. This is what happens for example in the case of insertions.

(23) a. I woke up (you're going to laugh //) in the morning at five o'clock // (constructed example)
b. { I studied | (sorry//) I studied in college | I studied } international relations // (ELUI)

We propose two equivalent ways to note this, either by placing the inserted utterance between parentheses as in (23) or by using the symbol # to indicate that the utterance is continued later at the following occurrence of #:

(24) a. I woke up #// you're going to laugh //# in the morning at five o'clock //
(constructed example)

These two notations are strict equivalents __"(" = "#//" and ")" = "//#"__, but the symbol # also allows the encoding of more complex cases such as the following example, where SPK1 is interrupted three times by SPK2. This does not keep SPK1 from pursuing a relatively complex utterance, all the while interacting with SPK2 through *yeahs* which punctuate SPK2's interventions. The sequence of //#+ tags indicates that the IU is completed (//), but that the DU continues later on (#+):

(25) SPK1: but but otherwise uh well & // in any case the fundamental research it it remains free //#+
SPK2: yeah yeah //
SPK1: #luckily //#+
SPK2 so yeah // in 2009 //
SPK1: yeah //
SPK2 : we'll have to see later //
SPK1: yeah // # the applied research < less // ^but the fundamental research < yeah //
(translation, *Rhapsodie*)

3.3 Embedded IUs

Direct discourse presents a particular difficulty due to the embedding of illocutionary acts. Consider the following example:

(26) he said [go away > poor fool //] //
(translation, radio)

The reported speech in (26), annotated with the symbols [ ], has its own illocutionary force, it can be regarded therefore as an autonomous IU. Regardless, the preceding segment (he said) does not form an autonomous illocutionary act or a complete DU. We treat such a structure as an embedded IU. The reported speech is an IU embedded in the IU made up of the whole utterance *he said go away poor fool*.

Another phenomenon that we treat as the embedding of IUs is the graft. We define a graft as the filling of a syntactic position with a segment belonging to an unexpected category (Deulofeu, 1999).

(27) a. you don't have an agenda with [one day I do this // one day I do that //] (translation, Deulofeu 1999)
b. you follow the tram line which passes towards the [I think it's an old firehouse //] // (translation, *Rhapsodie*)
c. I could like take and see {if I & | if it was worth it that I should go into "you know" more depth | ^or if that was just sort of like [okay {I l- | I like it} // ^but I don't wanna like study that // ^so I don't know //] } // (Micase)
d. we had criticized the newspaper [I think it was the Provencal #] we had criticized it in relation to (# or the Meridional //) in relation to the death of [what was his name // not Coluche // the other guy //] //
(translation, Blanche-Benveniste 1990)

This phenomenon can be regarded as a rupture of sub-categorization. The grafted segment usually has its own illocutionary force, being in most cases a unit commenting on the lexical choice that should have been done to respect the sub-categorization. In a graft, as well as in reported speech, an IU occupies a governed position inside a DU.

3.4 Associated IUs

A number of discourse particles (such as "right", "of course" in English, "quoi", "bon" in French) and parentheticals units (such as "I think", "I guess", "you know" in English, "je crois", "tu vois" in French) are endowed with an illocutionary force. However, these elements do not serve the purpose of modifying the common ground between speakers. They merely have a function of modal modification or interactional regulation. We call these units "associated units", we treat them as non autonomous illocutionary components and we annotate them between quotation marks " ".

(28) it's a really "you know" open field "you know" like all that stuff // (Micase)
(29) he is coming "I guess" // (constructed)
(30) "I mean" English wasn't that helpful itself // (ELUI)

4 Levels of annotation

Our annotation strategy rests on the fact that relatively good tools for automatic analysis of French written texts are currently in existence (Bourigault et al. 2005, De la Clergerie 2005, Boulier & Sagot 2005). Adapting these tools to spoken French would constitute a project in and of itself, one much more ambitious than our annotation project (even though we believe that *Rhapsodie* is an essential step towards the development of parsers for spoken language, and that one of the final uses of *Rhapsodie* will be as a

contribution to the training and development of these parsers). In other words, we want to use these tools developed for written text without modifying them substantially. In order to do this, we realize a pre-treatment of transcribed text "by hand": We manually annotate every phenomenon typical of the syntax of speech. The result is a pre-treated text that parsers can analyze as written text with minimal error. The segmentation into IUs and DUs described in the previous sections aims at providing such a pre-treatment. As we hope we have shown, our pre-treatment has a theoretical and practical value, and could constitute a satisfying analysis of speech on its own. Regardless, we would like to present all levels of our treatment, as this will allow a greater understanding of the choices that have been made (for example the analysis of piles during pre-treatment).

Our annotation procedure is organized into several steps which alternate regularly between automatic and manual treatment.

Level 1: Raw transcription (i.e. without syntactic enrichment) - This consists of orthographic transcription which includes speech overlap, truncated morphemes, etc.

Level 2: Simple automatic pre-treatment - Annotation of trivial "disfluencies" (such as word repetition) and identification of potential associated IUs (*um*, *uh*... but also *like*, *you know*...). This automatic step is very rough and is to be corrected at level 3.

Level 3: Manual syntactic segmentation - This is the annotation presented in the previous sections of this paper, indicating DUs, IUs, ICs, piles, etc. This level is obtained manually starting at level 2. The general idea is that it simultaneously constitutes:

- A coding of everything that we know we are not able to automatically calculate, and which would cause problems for parsers (originally programmed for written text),
- A coding which is satisfactory in itself and permits a preliminary study of the syntax-prosody interface.

A tool has been developed for checking the well-formedness of this level of annotation.

Level 4: Parser entry - Existing parsers for French have not been programmed to process simple transcriptions of speech, nor have they been tuned to treat the markup that we have introduced at level 3. However, these tags allow us to automatically segment the text and furnish the parser with sections it is capable of analyzing. The following example will illustrate this point.

(31) are you thinking {of other communicat~ | "uh" of other functions}
(constructed example)

would give us to two segments:

(32) a. are you thinking of other communicat~
b. are you thinking of other functions.

Certain fragments of text are therefore duplicated and analyzed multiple times. These analyses, if identical, are automatically fused in the ulterior levels. If they differ, a manual treatment is necessary. Another strategy consists of not unpiling but rather perceiving an utterance including a pile as a Directed Acrylic Graph (DAG), that is to say a graph in which the arcs are labeled by words of the text, and which integrate all possible paths in a pile structure. A parser like SxLFG (Boulier and Sagot 2005) can manage a DAG entry, but for the moment it is parameterized to choose the best path in the DAG and not to analyze the entire DAG.

Level 5: Parser output - Parsers provide us with a syntactic analysis in the form of a dependency tree. We now have two things left to do: 1) automatically translate these analyses so that they correspond exactly to the desired labels (this is mainly a renaming process of functional labels); 2) apply syntactic annotations computed for the unfolded segments to the original texts (those from level 3), while fusing duplicated syntactic annotations.

Level 6: Dependency analysis - This consists of level 5 after automatic reinsertion of analyzed sections and manual correction. The last level is a manual correction of level 5, this is absolutely necessary as the parsers still make many mistakes (we estimate that about 30% of dependencies will have to be corrected) and do not use our same labels. The encoding of level 6 is therefore a complete syntactic analysis of text, which includes microsyntax (functional dependencies) as well as macrosyntax.

Conclusion

The ongoing process of annotating transcriptions of spoken French with syntactic functions has revealed the necessity of a well-defined text segmentation separated into illocutionary and dependency units. This process is an interesting challenge in its own right as it allows, for the time being, only very limited automated steps,

and can be seen as a necessary pre-treatment before the parsing process, relying mainly on tools tuned to work on written data. Linguistically, contrary to the conventional ad-hoc punctuation of written text, our segmentation can be seen as a systematic punctuation process relying on reproducible criteria allowing for a distribution of this process to trained annotators. Moreover, the notion of paradigmatic piles naturally completes the short-comings of head-descriptions in coordinations and other paradigmatic phenomena.

If we want to share tools and resources across languages and theoretical models, it is necessary that annotation norms develop in the field of syntactic annotation of spoken texts, in other words, we need some kind of language-independent punctuation scheme reflecting syntactic and pragmatic segmentation of the text. This is a process that is well on its way for written text. Our work on French and English shows that our annotation scheme proposes criteria that can be applied to different languages while yielding interesting results. We hope this to be a contribution to the development of unified annotation methods in dependency annotation of spoken text and thus to a deeper understanding of functional syntax as a whole.

References

Abeillé, A., L. Clément, F. Toussenel. 2003. Building a Treebank for French. A. Abeillé (ed) *Treebanks*. Kluwer, Dordrecht.

Benveniste, E. 1966. *Problèmes de linguistique générale*, Gallimard, Paris

Berrendonner, A. 1990. "Pour une macro-syntaxe". *Travaux de linguistique* 21: 25-31.

Bertrand, R., P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, S. Rauzy. 2009. Le CID - Corpus of Interactional Data - Annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, 49(3): 1-30.

Bilger M., Blasco, M., Cappeau, P., Sabio, F. & Savelli, M.-J. 1997. Transcription de l'oral et interprétation: illustration de quelques difficultés. *Recherches sur le français parlé* 14: 55-85.

Blanche-Benveniste, C., M. Bilger, C. Rouget, K. van den Eynde. 1990. Le Français parlé. Études grammaticales. Paris, CNRS Éditions.

Blanche-Benveniste, C. 1997. *Approches de la langue parlée en français*, Ophrys, Paris.

Boullier, P. et B. Sagot. 2005. Analyse syntaxique profonde à grande échelle: SxLfg. *Traitement Automatique des Langues*, 46(2):65-89.

Bourigault D., Fabre C., Frérot C., Jacques M.-P. & Ozdowska S. 2005. Syntex, analyseur syntaxique de corpus, in Actes des 12èmes journées sur le *Traitement Automatique des Langues Naturelles*, Dourdan, France

Creissels D. 2004. *Cours de syntaxe générale*. Chapitre 1, http://lesla.univ-lyon2.fr/sites/lesla/IMG/pdf/doc-346.pdf

Culicover P., R. Jackendoff 2005. *Simpler Syntax*. Oxford: Oxford University Press

Cresti, E. 2000. *Corpus di italiano parlato*. Accademia della Crusca, Florence.

De la Clergerie, E. 2005. DyALog: a tabular logic programming based environment for NLP. *Proceedings of 2nd International Workshop on Constraint Solving and Language Processing*, Barcelona, Spain.

Degand, L., Simon, A. C. 2009. On identifying basic discourse units in speech: theoretical and empirical issues. *Discours* 4 (http://discours.revues.org/index.html)

Deulofeu, J. 1999. *Recherches sur les formes de la prédication dans les énoncés assertifs en français contemporain (le cas des énoncés introduits par le morphème que)*. Thèse d'état, Université Paris 3.

Dister, A., Degand, L., Simon, A. C. 2008. Approches syntaxiques en français parlé: vers la structuration en unités minimales du discours. *Proceedings of the 27th Conference on Lexis and Grammar*, L'Aquila, 10-13 September 2008, 27-34.

Gerdes, K., Kahane, S. 2009. Speaking in Piles. Paradigmatic Annotation of a Spoken French Corpus. *Proceedings of the fifth Corpus Linguistics Conference*, Liverpool.

Guénot M.-L. 2006. La coordination considérée comme un entassement paradigmatique: description, formalisation et intégration, *Proceedings of TALN*, Leuven, Belgique, 178-187.

Mel'cuk, I., Pertsov, N. 1987. *Surface Syntax of English. A Formal Model within the Meaning-Text Framework*, Benjamins, Amsterdam.

Searle, J. R. 1976. A classification of illocutionary acts. *Language in Society* 5:1, 1-23.

Simpson-Vlach, R., & Leicher, S. 2006. *The MICASE handbook*, The University of Michigan Press, Ann Arbor.