Speaking in piles: Paradigmatic annotation of French spoken corpus

Kim Gerdes, Sylvain Kahane

Speaking in Piles
Paradigmatic Annotation of a French Spoken Corpus

Kim Gerdes
LPP / ILPGA
Sorbonne Nouvelle & CNRS
kim@gerdes.fr

Sylvain Kahane
Modyco
Université Paris Ouest Nanterre La Défense & CNRS
sylvain@kahane.fr

Abstract

This article describes a central part of the syntactic schemes that are currently used in the ongoing annotation of a French spoken corpus. Based on the Aix School grid analysis of spoken French, the notion of « pile » is introduced, allowing for an elegant description of various paradigmatic phenomena like disfluency, reformulation, apposition, instanciation, including question-answering and colon effect, and different types of coordination. Piles naturally complete dependency annotations by modeling non-functional relations between phrases.

1. From Disfluency to Coordination

Disfluencies are a major aspect of spontaneous speech and a well-known difficulty for the annotation of Spoken Corpora (Shriberg 1994). We believe that it is important to include disfluencies in the higher level annotation schemes of spoken corpora, for theoretical as well as for practical reasons: Encoding disfluencies reveals basic aspects of utterance planning and constitute essential training data for future parsers of spontaneous speech.

In the Spoken Dutch Corpus, for example, only the repairs will be taken into account when constructing the syntactic structures. When complete constituents are repeated they will all be constructed up to that level, but only the last one will be part of the structure assigned to the utterance as a whole (Schuurman et al. 2004). Aside from the fact that the restriction to "corrected" sentences excludes large chunks of text from syntactic annotation, it is difficult to clearly determine the extension of disfluencies: Their start and end points are difficult to establish and, moreover, we will give examples in order to establish a continuum from disfluencies to standard coordinated structures, which makes it unnatural to include some and not others, as shown by Blanche-Benveniste (1990).

Sentence (1) is an example of a typical disfluency where the speaker hesitates many times and corrects herself. It is interesting to present such examples by aligning each part of the utterance where the speaker picks up again. The original English examples in this paper stem from the Micase corpus (Simpson-Vlach & Leicher

2006), in particular from the segment *Honors Advising*, Transcript ID: ADV700JU023.

(1) **a.** *mhm i wro- i w- i'm interested in the um international aspect more of um of a program or whatnot so like the international business i was gonna do*

b. i wro-
i w-
i'm interested in the "um" international aspect more of a "um"
of a program
or whatnot
so **like the international business i was gonna do**

This representation pattern, called a grid analysis, has been proposed by Blanche-Benveniste et al. (1979) and studied in numerous publications of the Aix team (for instance Blanche-Benveniste 1990, Bilger et al. 1997).

Shriberg (1994), following Levelt (1983), proposes to analyse disfluencies in three distinct segments: reparandum, (optional) interregnum, and repair. In (1), reparandums appear in normal type face and repairs are in bold face. Note how only few words (in bold face) would be annotated if the syntactical annotation were to be limited to repairs. This is even more apparent in the following slightly simplified example borrowed from Blanche-Benveniste:

(2) *donc pour essayer un petit peu de sortir cette personne de la misère*
(car c'est vraiment un petit peu semblable aux Misérables de Victor Hugo)
nous essayons tant bien que mal de lui faire comprendre ***que sa cabane***
dans quelques années (entre parenthèses elle a 79 ans)
quand elle aura des difficultés (ce qu'on espère pas)
des difficultés à se déplacer ou à évoluer
(c'est-à-dire qu'il y a énormément d'escaliers à monter pour arriver à sa cabane)
donc le jour où elle ne pourra plus se déplacer
ou qu'elle sera malade un petit peu plus sévèrement
on essaye de lui faire comprendre qu'elle ne pourra plus vivre dans cette cabane

so to try a little bit to get this person out of misery
(because it's really a bit similar to the Misérables by Victor Hugo)
we try the best we can to make her understand **that her shack**
in a few years (in parentheses she is 79 years old)
when she will have difficulties (which we don't hope for)
difficulties moving or turning around
(meaning that there are lots of staircases to go up in order to get to her shack)
so the day she won't be able to move around
or she will fall sick a little bit more seriously
we try to make her understand that she won't be able to live in this shack.

In this utterance, *que sa cabane* 'that her shack' (in bold) is the beginning of an embedded clause that never ends. This unsaturated first clause is reformulated into a new clause (***nous essayons*** *de tant bien que mal* ***de lui faire comprendre que sa cabane*** *…* → ***on essaye de lui faire comprendre qu'elle*** *ne pourra plus vivre dans cette* ***cabane***). The first clause cannot simply be considered as a reparandum and be skipped because it contains lots of very important adjuncts and parentheses (*dans quelques années, quand elle aura des difficultés, le jour où elle ne pourra plus se déplacer,* etc.*)* where the speaker elaborates point for point the reasons why the old lady should give up her shack. In this kind of list, it is difficult to keep voluntary and involuntary elaboration apart. Note that it is not only difficult to distinguish disfluencies from voluntary reformulation, but it is equally non trivial to tell reformulation apart from coordination. We will propose a unified description of these phenomena in section 3.

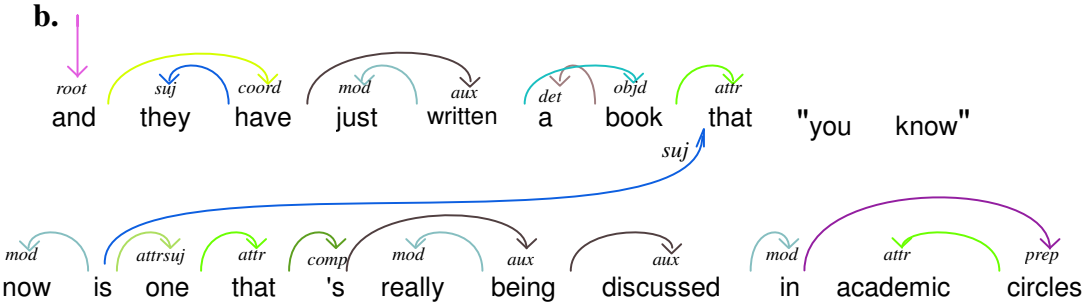
2. The Rhapsodie Project

These challenging data and original analyses are at the heart of the syntactic annotation of the Rhapsodie corpus. The ongoing 4-year project (2008-2011, http://rhapsodie.risc.cnrs.fr/en) directed by Anne Lacheret and sponsored by the French National Research Agency (ANR), consists of developing a 3 hours (36,000 words) corpus of spoken French with the key features of being free, representative, and annotated with multiple layers: All sound files, transcriptions, and annotations will be free resources in the sense that they can be downloaded, modified, and redistributed freely, at least for research purposes. The corpus is composed of various recent spontaneous spoken standard French sources. The sound and the orthographically corrected transcription are syllable aligned and will be annotated with phonological, prosodic, and syntactic layers of information.

We use a dependency type structure for the encoding of functional links. For utterances without paradigmatic piles, this boils down to an arguably classical dependency structure. Note though the double role that plays the relative pronoun in this analysis, resulting not in a simple tree but in a graph structure.

(3) **a.** *and they have just written a book that "you know" now is one that 's really being discussed in academic circles "um"*

b.



Note that discourse markers like *um* or *you know* are not considered as part of the dependency structure. They are surrounded by double quotes. This paper focuses on the innovative encoding of paradigmatic phenomena in the syntactic annotation.

Just like in an X-bar structure, functional dependency annotation supposes that a single word or morpheme can be identified as the head of any substructure. This view is challenged by many syntactic phenomena like for example coordination. The Prague Dependency Treebank (Hajič 1998) uses the coordinative conjunction (or even punctuation marks) as the head of the subtree, the Meaning-Text approach (Mel'čuk 1988) prefers the first conjunct to be the head, which subsequent conjuncts depend on. Similarly, some X-bar approaches use co-heads for coordinations in order to avoid the identification of a single head (Jackendoff 1977). Equally, the Negra-type annotation used in Alpino (van der Beek et al. 2002) and CGN (Dutch Spoken Corpus, Schuurman et al. 2004) falls back onto flat headless constituents for the description of coordinations. Contrarily to approaches that tend to encode paradigmatic and syntagmatic relations with the same tools, we adopt a two-dimensional analysis in two inter-related but different structures that will be presented in the following section.

French has surprisingly few syntactically annotated corpora, the most important project being the Paris 7 Treebank (Abeillé et al. 2003), a newspaper corpus of about one million words, annotated with Penn Treebank style phrase structure on the whole and syntactic functions on a quarter of the text, the whole corpus being under restrictive licensing.

The intensive study of the syntax of spoken French, notably at the University of Provence in Aix, looks back on more than 30 years of history, a starting point being the creation of the journal *Recherche Sur le Français Parlé* in 1977 (http://sites.univ-provence.fr/delic/rsfp/index.html). Nonetheless, the Rhapsodie corpus will be the first corpus of spoken French with coherent and complete syntactic annotations, and with the ambition to become a gold standard. Another large-coverage annotated corpus for spoken French is the Valibel corpus, a corpus of written and spoken Belgian French, which has been segmented into discourse units, that is more or less maximal domains with dependency relations (Degand & Simon 2005, Dister 2007). These two research groups' know-how, people and corpora, constitute the central basis for the Rhapsodie corpus construction and syntactic annotation.

The most comparable project to Rhapsodie in terms of annotation structures is CGN, a 10 million word transcribed spoken language resource, containing a subcorpus of one million words that is syntactically annotated with dependency relations. This is notably larger than the Rhapsodie project, which made the automation task a more important issue than for Rhapsodie. However, the syntactic annotation of CGN follows Negra (Brants et al. 2003) conventions combining constituency and syntactic functions into one annotated acyclic graph, allowing for multiple head analyses of coordinative structures. Disfluencies are excluded from the syntactic annotation as only the corrected sentence is included in the graphs.

3. Encoding of paradigmatic piles

3.1. Grid analysis and pile markers

We start our analysis of paradigmatic piles with simple examples which illustrate clearly the differences and similarities between paradigmatic phenomena like coordination or disfluency.

In the constructed examples (4a) below, we can perceive a reparation (i.e. that Felix may not be a linguist, like in example (4b) or (4c)) equally well as an additive coordination (i.e. Felix is not only a linguist but maybe also a computer specialist like in (4d)).

(4) **a.** *Felix is a linguist, maybe a computer scientist*

b. *Felix is a linguist uh maybe a computer scientist*

c. *Felix is a linguist or maybe a computer scientist*

d. *Felix is a linguist and maybe a computer scientist.*

Syntactically, (4c) is commonly considered as a coordination like (4d) although (4c) has the same interpretation as the clearly repaired sentence (4b).

For this reason, we want to give a unique syntactic structure to (4a) whatever the interpretation of the utterance may be. This structure should also resemble the structures of the 3 other example sentences and the grid analysis as proposed by Blanche-Benveniste et al. (1979) reveals this similarity.

We see in (5) that the grid analyses of the four sentences we propose are similar: We align segments, the *conjuncts*, that are in a paradigmatic relation (*a linguist* and *a computer scientist*), into different *layers*. In between these segments are words that we call *pile markers* (in italic).

(5) **a.** Felix is a linguist
maybe a computer scientist

b. Felix is a linguist
"uh" maybe a computer scientist

c. Felix is a linguist
or *maybe* a computer scientist

d. Felix is a linguist
and *maybe* a computer scientist

We have to distinguish three types of pile markers:

1. In (5d), pile markers like *and* or *or*, usually called *coordinating conjunctions*, are in a syntagmatic relation only with the conjuncts and do not play any role in the combination of the conjuncts with the context as they can only appear between two conjuncts: **Felix is and computer specialist.*
2. On the contrary, pile markers like *maybe, also, too,* or *not* are part of the syntagmatic relation between the second layer and the context: *Felix is [maybe/also/not] a computer scientist [too]*. In this sentence, *maybe, also, not* or *too* opens a potential list of elements that are in a paradigmatic relations with the following item (*a computer scientist*). For this reason, these words are referred to as *paradigmatising adverbs* (Nølke 1983, Masini & Pietrandrea 2010).
3. Some discourse markers like *uh* can play this paradigmatizing role as in (5b), but discourse markers are commonly not considered to be part of the syntactic structure, because they don't construct syntagmatic relations with other words.

In the structures (5) above, all pile markers are shown in italic, coordinating conjunctions are underlined, and discourse markers are between double quotes and are independent of the vertical alignment, as in example (3).

3.2. Junction Point and Backtracking Point

The grid analysis is not the only attempt to analyse disfluencies. Heeman, McMillin & Yaruss 2006 propose the following annotation of stammerers' productions:

(6) okay so what|what changed your mind

In (6), the disfluency is analysed as a case of backtracking. In some point (symbolized by a vertical line) the production is "interrupted" and the speaker goes back to a previous point (the head of the arrow). We propose the following notation:

- a pipe symbol ("|") for the *junction point*, i.e. the *interruption point* for disfluencies;
- an opening curly bracket ("{") for the *backtracking point*.

This notation is used in (7a). In (7b), we align the junction point | with the backtracking point { and we see that this new notation is more or less equivalent to the grid analysis.

(7) **a.** okay so { what | what changed your mind

b. okay so { what
| what changed your mind

The junction point and the backtracking point don't have exactly the same status. The first layer of a pile stands in an ordinary syntagmatic relation with its left context and the last layer of a pile can equally stand in a syntagmatic relation with its right context, and these borders are generally not linguistically marked. The left border of the first layer is sometimes marked either lexically (*Felix is neither linguist nor computer scientist*) or prosodically (for instance if the speaker wants to clearly mark the wide scope in *free software and corpora*). In the other cases, the backtracking point, is only identified *a posteriori* when the second layer starts at the junction point.

3.3. Parenthetisation

We can also see { as the beginning of a parenthetisation of the different layers. But the role and the position of the closing curly bracket ("}") is less clear. Let us consider (8a), an extended version of (6-7) with a third layer. This third layer is a group coordinated with the two previous layers and the beginning of the coordination is the same point as the backtracking point of the disfluency as shown in (8b):

(8) **a.** okay so what what changed your mind and what has it been changed to

b. okay so { what
| what changed your mind
| *and* what has it been changed to

c. okay so { what &
| what changed your mind
| *and* what has it been changed to }

d. okay so { { what
| what } changed your mind
| *and* what has it been changed to }

We can consider as in (8c) that *what* is corrected by the whole segment *what changed your mind* and that we have a unique pile with three layers. Or we could use embedded piles like in (8d), considering that only *what* was repeated the first time. If we adopt (8c) we describe the first *what* as a starting point of a syntactic structure that was never finished and we indicate that the first layer is unsaturated, using the ampersand (&) symbol.

Usually, prosody gives clear indications of the observed groupings: The analysis in (8c) represents an important break between the two *what*, whereas stuttering would be analysed as in (8d). In case of uncertainty because of a lack of clear prosodic evidence, our annotation rules prescribe (8d), where each layer in each pile finds its continuation in the right context and forms a saturated syntactic structure with it. Note that either choice of annotation allows the reconstruction of (8b), i.e. the analysis without final closing bracket, an analysis that could be preferred for the analysis of disfluencies (Heeman et al. 2006), to the somehow arbitrary choice of the closing bracket. Nevertheless, in the case of coordination, the closing curly bracket has a clearer role as it can indicate the scope of the coordination as for example in the analyses (9b) and (9c).

(9) **a.** a boy and a girl I met yesterday

b. {a boy | *and* a girl } I met yesterday

c. {a boy | *and* a girl I met yesterday }

We propose a graphical representation of our analysis inspired by both, the grid analysis and the representations of Heeman et al. 2006, where junctions are materialised by horizontal lines separating the conjuncts, which we call *junction bars* (Fig. 1). Left and right contexts, as well as the pile markers (in italics), are beside the junction bars. Hence the scope of the junction bars in the graphical representation is equivalent to the parenthetisation with bars and curly brackets.

left context	*pile markers*	conjuct$_1$	*pile markers*	
	pile markers	conjuct$_2$	*pile markers*	
	pile markers	conjuct$_3$	*pile markers*	right context

Figure 1. Graphical representation of a pile

Our two representations – parenthetisation and grid analysis with junction bars – can be compared in (10b) and (10c):

(10) **a.** and 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

b. and {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 }

c.

and	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

It follows from our definition that 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, for example:

- *and I have a lots of interests like "um" astronomy*
- *and I have a lots of interests that are a bit more like international relations*.

Note our analysis in (10) of the word *and*, which, in this sentence, does not act as a coordinating junction, i.e. as a marker in paradigmatic piling. As previously said, contrarily to paradigmatic adverbs, coordinating conjunctions only play a role in the syntagmatic relations between layers and are not parts of the paths. Each path can be analysed in terms of syntactic dependencies only, as all paradigmatic phenomena are by definition in the pile encoding.

3.4. Two-dimensionality of the Syntactic Structure

The different layers of a paradigmatic pile are first and foremost in a syntagmatic relation: Two (or more) layers combine together to form a pile. This is called the *junction*. The junction, though it is a syntagmatic operation, induces a paradigmatic relation between the conjuncts.

In this section, we would like to point to the fact that the dependency, i.e. the set of subcategorising and modifying relations, and the junction are two orthogonal modes of combination. The junction of the different layers gives a pile but the pile is not a phrase from the viewpoint of the dependency structure: A pile does not generally combine with its right and left contexts as a whole. We rather consider that the right

context combines with the first layer and the left context with the last layer and that the pile is invisible for the dependency structure. In particular, coordinating conjunctions have no syntactic position in the dependency structure: they are only part of the paradigmatic pile and mark the junction between the layers.
The relative independence of the two structures – the dependency structure and the junction structure – can be illustrated by (11).

(11) **a.** *Felix is neither a linguist nor a computer scientist.*

b. Felix is { *neither* a linguist | *nor* a computer scientist }

If we consider the combination of the pile with the left context (*Felix is*), we see that *neither* does not play any role and the syntactic head – i.e. the element controlling the distribution – of the phrase *neither a linguist* is *a linguist* and we will have a dependency between *is* and *linguist*. But if we consider the junction between the two layers, we see that *neither* and *nor* validate each other and must be considered as the heads of the layers. Moreover the syntagmatic relation between the two layers induces a third relation: the paradigmatic relation between the conjucts.

If, however, we restricted ourselves to a bracketing like phrase structure grammars do, we would be obliged to decide which dimension to favour. It is this dilemma that has caused the clash between those that consider conjuncts as co-heads (and thus favour the paradigmatic relation to the detriment of the link between the first conjunct and the left context, cf. Jackendoff 1977; (12a)) and those that consider the second conjunct as an adjunct (and thus give priority to the relationship between the first conjunct and the left context and the role of the coordinating conjunction in the combination of the two layers, cf. Steedman 1985, Mel'čuk 1988, Borsley 2005; (12b)).

(12) **a.** Felix is [[a linguist]$_{NP}$ and [a computer scientist]$_{NP}$]$_{NP}$

b. Felix is [a linguist [and [a computer scientist]$_{NP}$]$_{ConjP}$]$_{NP}$

The analysis (12b) is more easily tenable than (12a) but it does not shed any light on the paradigmatic relation between the conjuncts and it is difficult to see how it can be reasonably extended to obtain a suitable representation for (11b), i.e. where *neither* is analysed in the same way as *nor*.

The analysis of paradigmatic piles becomes even more complex in presence of paradigmatising adverbs. Consider:

(13) **a.** *Felix is a linguist and not a computer scientist.*

b. Felix is { a linguist | *and not* a computer scientist }

We would like our second path to be *Felix is not a computer scientist*. But in this last sentence, *not* is an adverb depending on the verb *is*. As the second layer is not related to the left context, how can *not* inherit its right syntactic position in the main sentence? We consider that a layer has a special position for the paradigmatising adverbs:

b. Felix is { { } {a linguist} | *and* {*not*} {a computer scientist} }

We propose the same analysis when a pardigmatising adverb is in the first layer:

(14) **a.** *Felix is not a linguist but a computer scientist.*

b. Felix is { {*not*}{a linguist} | *but* { } {a computer scientist} }

It follows from this proposition that the first layer receives two dependencies from the left context, which means that it is not a syntactic constituent in the dependency structure (a constituent is always the projection of one head). But the layer can be a constituent in our second dimension – the paradigmatic pile.

4. Pile Typology

Until now we have not established clear criteria about when we have encountered a pile and not a syntactic dependency. Our definition of the paradigmatic piles was only based on common properties of coordination and reformulation. We consider that a segment Y of an utterance piles up with a previous segment X if Y fills the same syntactic position as X. In other words Y does not occupy its own syntactic position in the utterance in terms of syntactic dependencies. Rather than being directly dependent on another term of the utterance, Y inherited its syntactic governor from its paradigmatic relation with X. Note that in fact Y is generally both in a syntagmatic and a paradigmatic relation with X: it is because Y combines in a syntagmatic way with X (or sometimes with a larger segment containing X) that it creates a paradigmatic relation and inherits a syntactic governor.

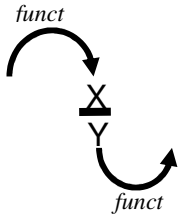


Figure 2.

Simplified pile schema with an incoming and an outgoing dependency link

We will now show that there are various other phenomena and not just coordination, disfluency, and reformulation that fall in the domain of paradigmatic piles and give a rough typology of these phenomena.

Note first that our definition of paradigmatic piles presupposes that the first layer of the pile occupies a syntactic position and is not syntactically independent like a "sentence". It is actually difficult to fix the upper limit for piles when the segments are no longer subcategorised, i.e. when we have piles of "sentences". In our corpus annotation, we nevertheless keep the pile encoding when we observe an important parallelism between the clauses, in both the syntactic construction and the lexical choices. This is the case in (15) where the parallelism marks the contrast and in (16) where the three last clauses are reformulations.

(15) vous étiez sténodactylo — 'you were a stenographer'
vous êtes euh directrice de l'Express — 'you are um the director of l'Express'

(16) et vous remontez euh jusqu'à la une grande place c'est la place de Verdun
là sur la place de Verdun il y a mh la préfecture voilà
comme repère il y a la préfecture
donc quand vous arrivez sur la place de Verdun la préfecture est sur votre droite

you go up um until the a big square that's the Verdun square
there on the Verdun square there is a um the prefecture that's it
as a landmark there is the prefecture
so when you arrive on the Verdun square the prefecture is on your right

We consider five principal types of piles: coordination, reformulation, apposition, instantiation, and intensification. We will not come back to coordination or reformulations, including both disfluencies and voluntary reformulations, all of which have already been presented in section 1.

Apposition, like in (17), is not very far from a reformulation, but, in this case, the speaker does not try to (better) reformulate the same description, but adds a different viewpoint on the same entity:

(17) *Peter, my best friend, lives in London.*

The example (18) illustrates a case of apposition with the resumption of a lexical element (here *route* 'road') which is stylistically strongly marked and constitutes a construction by itself (in this example, the speaker is a journalist and we may suppose that the text was prepared in advance).

(18) { deux petites phrases
| deux vraies options } qui dessinent { votre route
| une route qui témoigne { d'une certaine
| d'une bonne
| d'une très bonne } conduite

two little sentences
two real options that sketch your path
a path that shows a certain
a good
a very good conduct.

Instantiation is an important family of phenomena of paradigmatic piles. The first sub-type of instantiation we present is the colon effect. This construction, considered first by the Aix research group, has its name from its common transcription in writing using a colon.

(19) *I bought a lot of office supplies: paper, glue, and so on.*

(20) did a humanoid species spring up in various places
{ in the world
| *not just* in Africa
| *but also* in Asia
| *and maybe also* in southern Europe }

In these two examples of the colon effect, a general term (*office supplies* or *world*) is instantiated by a list of hyponyms. The first layer of a colon effect often contains an indefinite noun like *thing*. Instantiation differs from coordination or reformulation by the fact that it does not presuppose the contiguity of the two layers of the pile. In such a case we use the notation {…|} … {|…} to model the pile.

(21) **a.** je dirais que vous avez donné quelque chose de plus à la femme
des armes de persuasion

I would say that you gave something more to the woman: arms of persuasion.

b. je dirais que vous avez donné { quelque chose de plus |} à la femme {| des armes de persuasion}

(22) **a.** ben en fait il y a pas mal de choses qui rentrent en compte déjà euh l'ambiance du magasin

Well, in fact there are quite a few things that contribute already uh... the atmosphere in the shop …

b. ben en fait il y a {pas mal de choses |} qui rentrent en compte déjà "euh" {| l'ambiance du magasin}

One could hesitate to consider such instantiations as piles. And yet, although the two layers are not contiguous as in the other cases of piles, the instantiation does occupy

the same syntactic position (V*ous avez donné* ***des armes de persuasion*** *à la femme,* 'You gave arms of persuasion to the woman'). Moreover, it is clear that the instantiation phrase, although excluded from being attached to the rest of the utterance by any reasonable syntactic dependency relation, is not a completely syntactically autonomous discourse segment: Piling it up is the best means of linking it to the rest of the utterance and of indicating its role.

Another case of instantiation is illustrated by question-answering.

(23) **a.** S1: *When do you plan to come?*
S2: *Not today, maybe tomorrow.*

b.

	when	do you plan to come
not	today	
maybe	tomorrow	

c. { when |} do you plan to come {| *not* today | *maybe* tomorrow }

In question-answering, the pile is generally discontinuous, but again we observe a syntactic cohesion of the whole utterance (even if two speakers are involved): the answer is not a syntactically autonomous discourse segment and it can be substituted for the interrogative pronoun (When *do you plan to* come → You *do* not *plan to come* today).

The last type of paradigmatic piles we want to present is rather different. In this case the repetition of a word or a phrase is used to produce intensification:

(24) **a.** *This is a very very serious question.*

b. *Make it quickly quickly quickly.*

c. *I gave examples examples examples.*

There are probably other types of piles. For instance, frozen constructions such as *the more you explain, the less they understand* could be other case of piling. In the current first annotation experiences in the Rhapsodie project, we use the following pile type hierarchy:

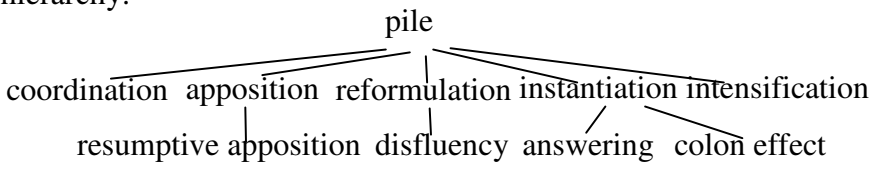


Figure 2. Hierarchy of pile types

Each junction will be typed, which boils down to indicating the nature of the relation between two contiguous layers. Junctions are also used to express an eventual bracketing of the pile: The number of lines indicates the importance of a junction.

(25) **a.** *We are looking for someone who speaks Italian or French and German.*

b.

	Italian
or	French
and	German

{Italian or French} and German

c.

	Italian
or	French
and	German

Italian or {French and German}

This encoding seems particularly suitable to the study of the syntax-prosody congruency, as we expect to find correlations between the importance of the junctions and the importance of the prosodic breaks.

5. Implementation

For the project, we have syntactic dependency parsers at our disposal with good results on written French. However, these parsers are unable to recognize most of the phenomena of paradigmatic piles in a satisfying manner. For this reason, we have decided to manually annotate paradigmatic piles. This allows for a "reconstruction" of "sentences" and to give to our parsers segments that they can analyse correctly. The "piling" process is the first step in the syntactic annotation procedure in the Rhapsodie project. It is entirely done manually (although further extensions of the syntactic annotation to new corpora may use machine learning procedures on the gold standard to be created). The 'piled up' corpus will be given to the automatic dependency parsers, which have not been specifically tuned for spoken texts and which would choke on punctuation-free transcriptions.
In order to pool the piling annotation process, a free web tool is under development that allows for different researchers to graphically pile the raw corpora and to classify the observed junctions. The program runs in Javascript (with underlying Python scripting on SQLite on a server) and uses SVG for the graphic rendering. The same tool will eventually be used for the manual correction process of the automatic dependency annotation.

6. Complex Cases and Further Investigations

In this section, we will point out some difficulties and surprising features of the pile encoding. We will give two examples of problems that arose when annotating the corpus. Both are due to the crossing of piles with other discourse phenomena: The first case concerns parentheses, the second "grafts". The following utterance (from an interview of Roland Barthes) contains a parenthetical phrase (in between brackets).

(26) *c'est euh au fond euh j'ai tendance (**et** c'est peut-être un peu ambigu) **mais** j'ai tendance à à penser par phrases disons et non pas à penser par pensées*

it's uh basically uh I tend to (**and** that's maybe a little ambiguous) **but** I tend to to think by sentences, let's say, and not to think by thoughts.

This case is difficult to encode. First, this parenthetical is introduced by *et* 'and'. In spite of this element which normally marks a coordination, we cannot call this construction a coordination, because the second layer is a metalinguistic commentary on the first layer, and these layers cannot commute in any way: (**c'est peut-être ambigu à penser par phrases* 'it may be ambiguous to think in sentences'). The central difficulty stems from the resumption *mais j'ai tendance* 'but I tend to' that is equally introduced by a typical coordination marker. While this second *j'ai tendance* 'I tend to' piles clearly with the first one (we have a case of "disfluency" here), the *mais* 'but' does not mark this pile (**j'ai tendance mais j'ai tendance* 'I tend to but I tend to'), but it marks the coordination with the parenthetical. A solution, not yet formalized, would be to introduce a third dimension to our two-dimensional piles in which the parentheticals would be represented; the whole would put the three segments in a triangular relation.

The next phenomenon we want to present is not directly related to paradigmatic piles, but it can interfere with them and give very complex productions. This phenomenon has been called *graft* by (Deulofeu 1999): The speaker does not find an adequate term in the lexicon and he produces a whole clause in a syntactic position

where a noun phrase is expected. In the next utterances, the graft is indicated by square brackets: […].

(27) **a.** it's being done in the whole field of
[{ is there a common ancestor
| *or* 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 }]

b. and i know they offer like
{ a w-
| half a term class
| *or* something {that if i had space &
| that 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] } } }

We see in these examples that a graft can be relatively elaborated with several paradigmatic piles (like in (a)) or even several "sentences" (separated in (b) by the delimiter //).

The following example from (Blanche-Benveniste 1990:151) illustrates the second type of problems we encountered:

(28) **a.** *on avait critiqué le le journal de je crois que c'était le Provençal on l'avait critiqué par rapport à ou le Méridional par rapport à la mort de comment il s'appelle pas Coluche l'autre*

one criticized the the newspaper of I think it was the Provençal one criticized it in relation to or the Méridional in relation to the death of what was his name not Coluche the other one

The problem stems from a graft. Two piles are added onto this structure: One pile is added into the matrix discourse (resumption of *on avait critiqué* 'one criticized') and another pile into the graft (the coordination of *le Provençal ou le Méridional* 'the Provençal or the Méridional') and the two intertwine. Moreover the speaker produces a second graft while searching for the complement of *la mort de* 'the death of'.

b. on avait critiqué le
le journal de
[je crois que c'était **le Provençal**
on l'avait critiqué
ou le Méridional] par rapport à
par rapport à la mort de [**comment** il s'appelle
pas Coluche
l'autre]

c. { on avait critiqué { {le | le} journal de & | [je crois que c'était {$_1$ le Provençal |}$_1$] | on l'avait critiqué} { par rapport à ({|$_1$ *ou* le Méridional }$_1$) | par rapport à} la mort de [{comment |} il s'appelle {| *pas* Coluche | l'autre}]

Due to the intertwining of two piles we prefer in our encoding to add a special mark (here the subscript 1) to the discontinuous pile that begins in the graft and reopens in a parenthesis later.

Our final examples show that even for the well-known challenging problem of non-constituent coordination, the pile encoding we propose allows us to obtain a relatively simple representation, coherent with our preceding analysis of spoken

language phenomena. For (29), we use a two-colons pile: each layer contains two syntactic positions (on top of the position of paradigmatic adverbs).

(29) **a.** *He gave Peter a book and John a disk.*

b. He gave Peter | a book
and | John | a disk

c. He gave { { Peter }{ a book } | *and* { John }{ a disk } }

Once again we see that our two-dimensional representation could not be easily reduced to a one-dimensional representation. Indeed, in the dependency structure, *Peter* and *a book* occupy two separate syntactic positions and do not form together a syntactic constituent. But in the pile, they are part of the same layer and form together a sort of syntactic constituent, as well as *John* and *a disk*.

This representation even permits us to encode question-answer situations with new topics. It allows for a complete analysis of these phenomena that are usually considered as syntactically deficient (see a similar only partially analysed example in the CGN annotation guide, Hoekstra et al. 2003). Note that the left colon of the pile plays a structural role comparable to paradigmatising adverbs.

(30) **a.** S1: *Which newspaper do you read?*
S2: *In the morning the Guardian, in the evening the New York Times.*

b.

	Which newspaper	do you read?
in the morning	the Guardian	
in the evening	the New York Times	

c. { { }{ which newspaper } |} do you read {| {in the morning }{ the Guardian } | { in the evening }{ the New York Times } }

Note that we have introduced an empty syntactic position in the first layer. We need it because *in the morning* and *in the evening* will not directly receive a syntactic function from the verb *read*: they receive a syntactic position by the fact that they pile up with this empty position which depends on *read*, the same position that *in the morning* would receive in the sentence *I read the Guardian in the morning*.

7. Conclusion

The formal mechanism of paradigmatic piles allows us to model phenomena that are difficult to describe in terms of syntactic dependency. Departing from the grid analysis, which was introduced by the Aix research group simply for facilitating the comprehension of the spontaneous speech transcriptions, we propose a first formalization of this annotation procedure. We showed several advantages and hitches of the pile annotation.

The realization of a pile annotated corpus requires a very precise definition of the extension of the covered phenomena. This includes the handling of the pile structuring elements, which brought us to the distinction of two types of pile markers. Note that our definition generalizes the binary *repair-reparandum* opposition (1st layer reparandum, 2nd layer repair) to piles that are lists of arbitrary length; and we do not suppose that an earlier layer is necessarily replaced by a subsequent layer. All layers can participate in the construction of the final meaning of the utterance. Equally, the pile markers generalize interregnums, because the former include coordinative conjunctions and adverbs. The pile annotation opens new views on coordination in

formal syntactic models by proposing a global approach to paradigmatic phenomena (cf. Guénot 2006 for a first attempt in this direction).

The realization of the Rhapsodie corpus is an ongoing project and by means of systematic annotations with paradigmatic piles, we expect to discover new aspects of spoken French, in particular concerning the relation between syntax and prosody. The corpus may also be useful for the development of innovative grammars and parsers that include paradigmatic phenomena from disfluency to non-constituent coordination.

Acknowledgements

We want to thank all the participants of the syntax group of the Rhapsodie project, with whom we had very fruitful discussions in our regular meetings and who played a fundamental role in the maturation of this work: Christophe Benzitoun, Jeanne-Marie Debaisieux, Anne Dister, Renaud Marlet, Frédéric Sabio, and Bernard Victorri. We also thank Anne Lacheret who urged us to work on these extraordinary data. We had many interesting exchanges with Paola Pietrandrea on the typology of piles. We are deeply indebted to Claire Blanche-Benveniste and José Deulofeu for very interesting exchanges of ideas, in particular in Claire's garden during our unforgettable stay in Aix in July 2008.

References

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

Beek L. van der, G. Bouma, R. Malouf, G. van Noord (2002), The Alpino dependency treebank. In M. Theune, A. Nijholt, H. Hondorp (eds.), *CLIN 2001. Selected Papers from the Twelfth CLIN Meeting*, Amsterdam.

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

Bonvino E., Masini F., Pietrandrea P. (2009), List Constructions: a semantic network, In *Proceedings of Aflico*, Paris.

Blanche-Benveniste C. (1990), *Le Français Parlé: Etudes Grammaticales*, CNRS, Paris.

Blanche-Benveniste C., Borel B., Deulofeu J., Durand J., Giacomi A., Loufrani C., Meziane B., Pazery N. (1979), Des grilles pour le français parlé, *Recherches sur le français parlé*, *2*, 163-205.

Borsley R. D. (2005), Against ConjP, *Lingua* 115, 461-482.

Brants T., W. Skut, H. Uszkoreit (2003), Syntactic Annotation of a German Newspaper Corpus, in A. Abeillé (ed.) *Treebanks: building and using parsed corpora*, Kluwer, Dordrecht, 73-87.

Degand L., A.C. Simon (2005), Minimal Discourse Units: Can we define them, and why should we?, *Proceedings of SEM-05*, Biarritz, 65-74.

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, University Paris 3.

Dister A. (2007), *Étiquetage morphosyntaxique de corpus textuels oraux. Le cas de la banque de données VALIBEL*, PhD Thesis, Catholic University of Louvain.

Gerdes K., S. Kahane (2007), Phrasing it differently, in L. Wanner (ed.), *Selected lexical and grammatical issues in the Meaning-Text Theory*, Benjamins, 297-335.

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

Hajič J. (1998), Building a syntactically annotated corpus: The Prague Dependency Treebank. *Issues of Valency and Meaning*, Karolinum, 106-132.

Heeman P., A. McMillin, J. S. Yaruss (2006) An annotation scheme for complex disfluencies. In *Proceedings of the 9th International Conference on Spoken Language Processing*, Pittsburgh.

Hoekstra H., M. Moortgat, B. Renmans, M. Schouppe, I. Schuurman, and T. Van der Wouden (2003), *CGN Syntactische Annotatie*.

http://lands.let.kun.nl/cgn/doc_Dutch/topics/version_1.0/annot/syntax/syn_prot.pdf

Jackendoff R. S. (1977) *X' Syntax: A Study of Phrase Structure*, Linguistic Inquiry Monograph, MIT Press.

Levelt W. J. M. (1989). *Speaking: from intention to articulation*. MIT Press.

Masini F., P. Pietrandrea. (2010) Magari, *Cognitive Linguistics, 21:1*, 75-121.

Mel'čuk I. (1988), *Dependency Syntax: Theory and Practice*, State Univ. of New York Press, Albany.

Nølke H. (1983), *Les adverbes paradigmatisants : fonction et analyse*, Akademisk Forlag, Copenhagen.

Schuurman I., W. Goedertier, H. Hoekstra H., N. Oostdijk, R. Piepenbrock, M. Schouppe (2004), Linguistic annotation of the Spoken Dutch Corpus: If we had to do it all over again ..., in *Proceedings of LREC*, Lisbon, 57-60.

Shriberg E. (1994). Preliminaries to a Theory of Speech Disfluencies, Phd Thesis, Berkeley University.

Simpson-Vlach R. C., S. Leicher (2006), *The Micase Handbook: A Resource for Users of the Michigan Corpus of Academic Spoken English*, The University of Michigan Press.

Steedman M. J. (1985) Dependency and Coordination in the Grammar of Dutch and English, *Language*, 61:3, 525-568.