Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie

Rachel Bawden[✧], **Marie-Amélie Botalla**[✧], **Kim Gerdes**[✦], **Sylvain Kahane**[✧]
[✧]Modyco, Université Paris Ouest Nanterre & CNRS
[✦]LPP, Université Sorbonne Nouvelle & CNRS
Email: rachel.bawden@keble.oxon.org, marieamelie.botalla@gmail.com,
kim@gerdes.fr sylvain@kahane.fr

Abstract

This article presents the methods, results, and precision of the syntactic annotation process of the Rhapsodie Treebank of spoken French. The Rhapsodie Treebank is an 33,000 word corpus annotated for prosody and syntax, licensed in its entirety under Creative Commons. The syntactic annotation contains two levels: a macro-syntactic level, containing a segmentation into illocutionary units (including discourse markers, parentheses …) and a micro-syntactic level including dependency relations and various paradigmatic structures, called pile constructions, the latter being particularly frequent and diverse in spoken language. The micro-syntactic annotation process, presented in this paper, includes a semi-automatic preparation of the transcription, the application of a syntactic dependency parser, transcoding of the parsing results to the Rhapsodie annotation scheme, manual correction by multiple annotators followed by a validation process, and finally the application of coherence rules that check common errors. The good inter-annotator agreement scores are presented and analyzed in greater detail. The article also includes the list of functions used in the dependency annotation and for the distinction of various pile constructions and presents the ideas underlying these choices.

Keywords: syntax, dependency, annotation, treebank, spoken language, French, inter-annotator agreement, paradigmatic constructions, pile constructions, coordination, micro-syntax, macro-syntax, semi-automatic annotation process, online annotation tools, Creative Commons, open data

1. Introduction

This article presents the methods, results and precision of the syntactic annotation process of the Rhapsodie Treebank. The Rhapsodie Treebank of spoken French is an open-source 33,000 word corpus annotated for prosody and syntax (Lacheret et al. 2011, Lacheret et al. 2014). The syntactic annotation contains two levels: a macro-syntactic level, containing a segmentation into illocutionary units (including discourse markers, parentheses …) (Deulofeu et al. 2011) and a micro-syntactic level, which is presented here.

The micro-syntactic annotation is comparable to major dependency treebanks. Its main originality lies in a rich annotation of paradigmatic phenomena called "piles" (Gerdes & Kahane 2009, Kahane 2012), integrated into the syntactic annotation scheme, including a syntactic and semantic classification (Kahane & Pietrandrea 2012). The high frequency of disfluencies and reformulations encountered in spoken language made the development of a general scheme that includes all paradigmatic phenomena a crucial step of the whole annotation process.

The micro-syntactic encoding involves the assignment of morphosyntactic information to each node (part of speech, lemma, plus other relevant features such as mood, tense, person etc.) and a micro-syntactic dependency analysis. The guide, developed prior to annotation and elaborated a posteriori to accommodate oversights is available in a French and an English version at www.projet-rhapsodie.fr, alongside different formats of the Rhapsodie Treebank itself.

In spite of good inter-annotator agreement, we applied hand-written rules to check the coherence of the resulting structures.

2. Workflow

The transcriptions of each recording underwent an initial segmentation according to lexematic word boundaries. A macro-syntactic analysis was produced for each sentence, describing the illocutionary groupings and providing clues for the micro-syntactic analysis.

The macro-syntactic annotation provided the segments of texts that were then analyzed by an automatic parser, FRMG, developed by Villemonte de la Clergerie (2010). The output of the parser was semi-automatically converted into the desired dependency relations and put into the collaborative online dependency annotation tool Arborator (Gerdes 2013). Using this tool, the sentences were analyzed by an average of two annotators, and then validated by a third annotator, who had access to the two previous annotations. The annotators had at their disposition the macro-syntactic encoding of each sentence, the annotation guides for both macro and micro-syntactic encoding, and the original sound recordings from which the transcriptions had been made.

Validation was followed by a final correction stage, whereby all validated trees were checked for correctness and consistency of analysis. This task will be discussed in the final section.

3. Dependency analysis

Our analysis of government relations mainly follows the tradition in dependency syntax (Tesnière 1959, Mel'čuk 1988, Kahane 2001). However we choose to encode fewer relations than the 30-odd relations used in some of the dependency annotations of the CoNLL shared task treebanks (Nivre et al. 2007, Johansson 2008), some of which are redundant and can be retrieved in combination with the POS annotation and other lexical features of the annotation. Traditionally, the set of functional relations used to label dependents contains a number of specific labels that distinguish categories (Mel'čuk & Pertsov 1987). This redundancy can be overcome by a complete inclusion of the dependent category in the function label as is the case in the Stanford annotation scheme: *det* (determiner), *amod* (adjectival modifier of the noun), *prep* (prepositional modifier), … (De Marneffe & Manning 2008). We take the opposite stance and limit our functional labels to encoding only relational rather than categorical differences, thus reducing our set of functional labels to those that are complementary to categorical distinctions. For example, we distinguish between the subject and the object for the nominal dependents of a verb. We refer to only seven plain dependency relations:

root: elements that are not governed by another element
sub: grammatical subjects of verbs
obj: direct objects of verbs.
obl: oblique complements of verbs, including indirect objects
ad: adjuncts to the verb
pred: all elements that form a complex predicate with a verb (past participles, verbal complements of modals, predicative adjectives …).
dep: all the dependents of non-verbal forms
junc: to link elements to junctors (= coordinating conjunctions), giving the asymmetrical annotation of coordination put forward by Mel'čuk (1988).

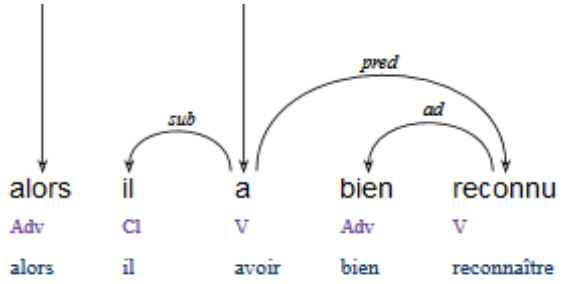

Figure 1: 'so he easily recognized (them)'.

4. Pile constructions

A particularity of the Rhapsodie project is the encoding of *paradigmatic relations* as 'piles', i.e. the identification of syntactic constructions based around elements that occupy the same syntactic position in the utterance (Gerdes & Kahane 2009, Kahane 2012). This is notably the case for coordinating constructions such as *I have an X and a Y*, where *X* and *Y* are in a paradigmatic relation as they are both in the same syntactic position governed by the verb form *have*. We say that *an X and a Y* form a *pile*. The paradigmatic relation between *X* and *Y* is encoded by a directed link labelled *para_coord* from *X* to *Y*.

We extended this analysis of coordination to other similar constructions common to spontaneous speech, in which two or more elements occupy the same syntactic position. Only a few attempts have been made in the past to syntactically analyse these constructions, and, to our knowledge, they have not been annotated in a dependency framework before. The Dutch Spoken Corpus (CGN, Hoekstra et al. 2003), for example, skips these constructions in order to obtain a classical dependency backbone. Each paradigmatic relation is assigned a type, mainly according to its semantic properties. We consider seven types of paradigmatic relations:

coordination: between coordinated elements (*para_coord*).
intensification: between elements repeated for intensification purposes as in ***des dizaines et des dizaines*** *d'années* ('**dozens and dozens** of years') (*para_intens*).
disfluency: between elements repeated due to hesitation in the formulation of the utterance, characterized by the repetition or partial repetition of a single lexeme as in *c'était un* ***un un un*** *enfin une super expérience* ('it was **a a a** well a super experience') (*para_disfl*).
double formulation: between elements which have the same denotation and can be substituted one for the other. We include in this definition qu-questions and answers if the answer occupies the same syntactic position as the interrogative pronoun in the question. Double formulation is an intentional speech act, and a way of testing whether a repetition belongs to this category is to prefix the second element by *c'est-à-dire* ('that is to say') (*para_dform*).
reformulation: between elements which have the same denotation, due to a speaker reformulating the utterance. A way of testing whether a repetition belongs to this category is to prefix the second element by *je veux dire* ('I mean') (*para_reform*).
hypernym: between elements which combine to form a hypernym, so that each element forms a subset of the denotation of the pile relation. Hypernymic relations are often

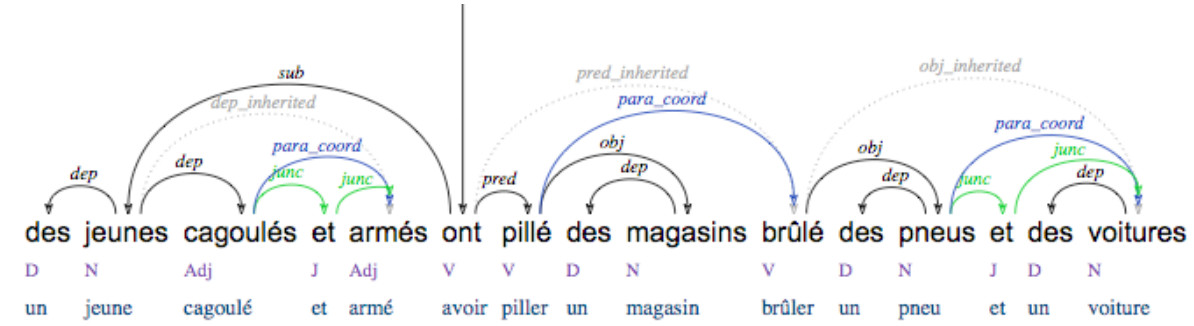

Figure 2: ‘hooded and armed youths looted shops, burned tires and cars’

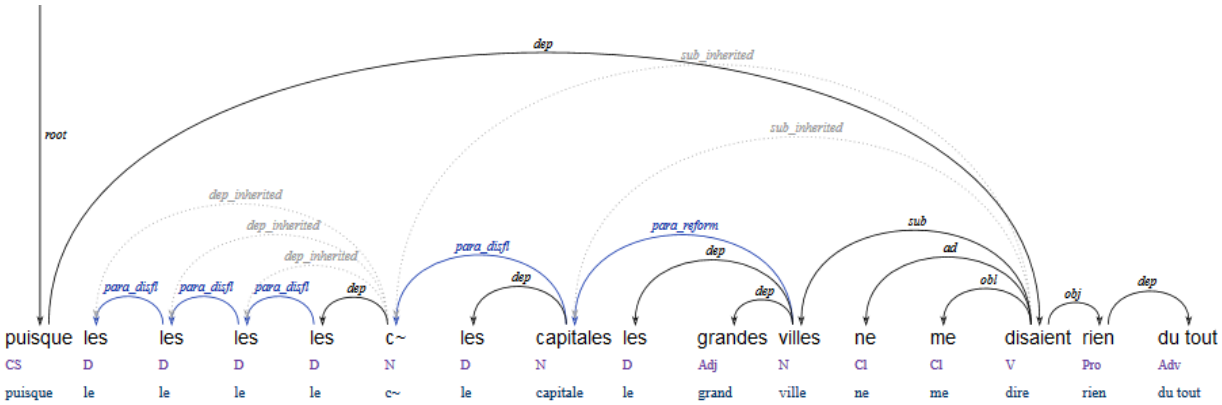

Figure 3: ‘since the the the the c~ capitals the large towns didn’t appeal to me in the slightest’

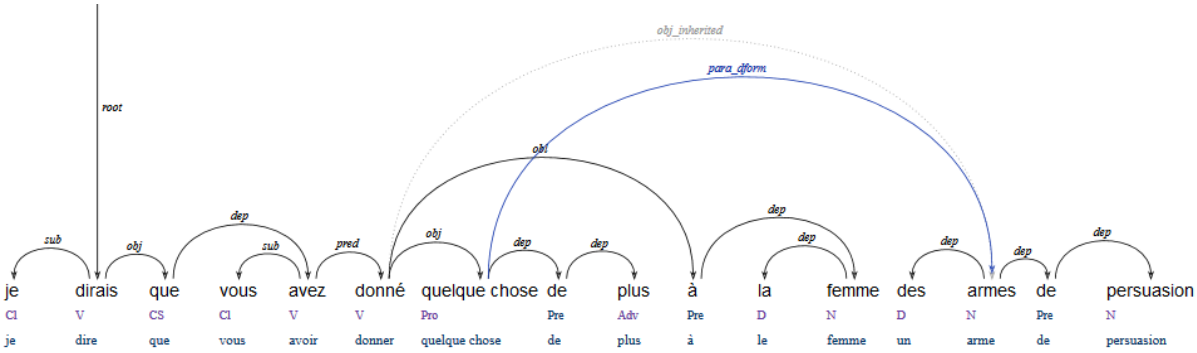

Figure 4: ‘I’d say that you have given something extra to the woman, weapons of persuasion’

present with general extenders such as *et tout ça* (‘and all that’), *et caetera*, which denote the rest of a subset (*para_hyper*).

negotiation: between elements indicating a request for confirmation, a confirmation, a refutation, or a correction. The word *enfin* (“well/after all/still”) is often an indicator of refutation as in *des des Français* ***enfin des Français***... (‘the French **well the French**’) (*para_negot*)

The presence of a paradigmatic relation also implies[1] the presence of an inherited dependency relation between the governor and the second conjunct (in grey in the figures). Since the two conjuncts have the same function in the

1 A notable exception is the case of general extenders such as *et caetera* and *et tout ça* ‘and all that’, which do not inherit a dependency, although they are considered to be part of a pile construction.

	dep	root	sub	ad	pred	obj	obl	junc	total
plain	14145	6162	4045	2675	2159	2115	927	916	33144
inherited	1284	324	219	315	266	316	135	44	2903
total	15429	6486	4264	2990	2425	2431	1062	960	36047

Distribution of the government relations in the Rhapsodie corpus

	disfl	coord	reform	dform	intens	negot	hyper	total
para_	831	557	260	194	126	118	77	2163

Distribution of the paradigmatic relations in the Rhapsodie corpus

	N	V	Cl	D	Pre	Adv	I	Adj	J	Qu	CS	Pro	Pre+D	X	Pre+Qu	total
POS	6249	5969	4177	4081	3457	2784	1978	1610	1141	800	726	718	484	198	3	34375

Distribution of the parts of speech in the Rhapsodie corpus

Figure 5: Numbers of dependencies and categories

pile, all the conjuncts of a same pile have the same governor and have the same type of dependency link, whether plain or inherited. The list of inherited relations is identical to that of the plain relations, except that each type is affixed as *_inherited* (See Figures 2 and 3).

Note that a pile can be discontinuous (*quelque chose de plus ... des armes de persuasion* 'something extra ... weapons of persuasion'), the corresponding paradigmatic relation being non-projective (the *para_dform* link crosses the *obl* dependency), see Figure 4.

In this example, the NP *des armes de persuasion* 'weapons of persuasion' forms an autonomous illocutionary unit, instantiating the indefinite NP *quelque chose de plus* 'something extra'. This configuration is similar to a wh-question followed by the syntactically incomplete unit that is its answer, as in "What do you give to the woman? – Weapons of persuasion."

5. Distribution of dependency relations in the treebank

The tables above illustrate the distribution of the dependency relations by type of relation. The numbers are based on the final corpus following validation of all annotated trees.

6. Agreement Analysis

Of the six annotators, three were considered expert, in that they participated in the initial elaboration of the dependency analysis used in annotation. According to Landis and Koch (1977)'s interpretation of the kappa coefficient, the agreement between the majority of the annotator pairs is considered almost perfect (inter-annotator agreement is between 0.76 and 0.95 with an average of 0.81).

As expected, the least disputed relation was that of the subject: for 95.82% of the cases where an annotator considers a *sub* dependency the other annotator does the same. The percentage of agreement is particularly low for the oblique complement: only 41.76% of the *obl* links are also annotated *obl* by the other annotator and 31.49% of them are annotated as adjuncts (*ad)* by the other annotator, translating the difficulty in confining the scope of verbal valency.

It is also noteworthy that paradigmatic relations were not confused with government relations, but 30.16% were forgotten or attributed to another conjunct by the other annotator.

The following matrix shows the distribution of an annotator's choice of label for a given paradigmatic link in relation to the label assigned by a second annotator for the same link, if the link was annotated as a paradigmatic relation.

The least disputed type is *para_disfl*, at 78.24% agreement. It is also the most common type of paradigmatic link in the corpus, making up almost 40% of the total pile links. The second least disputed label is *para_intens* (representing just 6% of the total links), followed by para*_coord*, which is the second most frequent label at just below 24% of the total links. However these three types are the only types for which agreement exceeds 50%. The percentage agreement for *para_hyper, para_reform*, *para_dform* and *para_negot* is significantly lower and dips as low as 10.14% for the label *para_negot*. Whereas a logical coordination of two different elements (*para_coord*), the repetition of the same element for intensification (*para_intens)* and the repetition of the same element due to hesitation *(para_disfl*) are relatively easy to identify, the other labels are more subjective, and the boundaries between each type more blurred. The line is notably blurred between *para_reform* and *para_dform*, where agreement was just under 30% for each label. The very low percentage scores for these four types of pile relation suggest that improvements need to be made in the types of pile relations; either in terms of the distinctions themselves or in terms of the indications given in the annotation guide, which should include more easily applicable tests.

7. Post-validation correction

According to the calculation of inter-annotator agreement,

	dep	sub	pred	obj	obl	ad	root	junc	para	none	total
dep	*81.30*	0.65	0.45	1.81	0.97	3.13	0.15	0.36	0.15	11.03	100
sub	1.85	*95.82*	0.15	0.06	0.03	0.09	0.00	0.00	0.00	2.00	100
pred	1.77	0.44	*84.28*	4.67	2.06	3.49	0.00	0.00	0.10	3.19	100
obj	7.22	0.14	6.13	*68.74*	5.37	2.38	0.05	0.05	0.09	9.83	100
obl	6.44	0.16	5.22	5.71	*41.76*	31.49	0.00	0.00	0.08	9.14	100
ad	8.59	0.11	1.56	1.15	8.89	*64.63*	0.37	0.11	0.00	14.59	100
root	0.54	0.00	0.00	0.04	0.00	0.24	*88.86*	0.00	0.00	10.32	100
junc	4.12	0.00	0.00	0.13	0.00	0.66	0.00	*77.82*	0.80	16.47	100
para	0.68	0.00	0.17	0.06	0.06	0.00	0.00	0.28	*68.59*	30.16	100
none	29.40	1.82	3.05	4.70	2.25	18.97	11.30	4.97	26.54	*0.00*	100

Distribution of dependency relation labels in comparison to a second annotator's choice (in %)

	_coord	_hyper	_intens	_disfl	_reform	_dform	_negot	none	total
_coord	*72.29*	6.69	0.64	0.32	5.10	8.92	2.86	3.18	100
_hyper	31.34	*14.93*	2.99	1.49	20.90	17.91	2.98	7.46	100
_intens	2.53	2.53	*65.82*	21.52	0.00	1.27	2.53	3.80	100
_disfl	0.18	0.18	3.05	*78.24*	15.11	0.72	1.08	1.44	100
_reform	5.93	5.19	0.00	31.11	*33.33*	14.44	7.78	2.22	100
_dform	18.79	8.05	0.67	2.69	26.17	*32.89*	4.70	6.04	100
_negot	16.07	3.57	3.57	10.72	37.50	12.50	*12.50*	3.57	100
none	23.26	11.63	6.98	18.60	13.95	20.93	4.65	*0.00*	100

Distribution of dependency relation labels in comparison to a second annotator's choice (in %)

Figure 6: Differences in annotations

based on the distribution and labeling of dependency relations, the level of agreement was high, showing that in most cases the annotation guide had been well followed and that there was relatively little disagreement amongst the analyses. However this calculation does not take into account errors made consistently by all annotators. Despite having two annotators plus a validator for the annotation of the treebank, human error remained relatively high, in particular for the direction of paradigmatic links, inherited dependencies, and parts of speech. In order to remedy this problem, we added an additional step to the annotation procedure.

For this last step of the workflow, we developed rules to determine the well-formedness of the trees, and searched for structures that did not obey them. In total we had 16 rules, most of which checked the compatibility between the dependency and the parts of speech of the words concerned. For example, a determiner is necessarily governed by a noun. The examples were automatically detected and manually corrected.

Type of error	Number of nodes corrected
More than one paradigmatic link to a single node	4
Paradigmatic link without inherited dependency	113
Paradigmatic link to a node which also has a plain dependency	49
Different types of plain/inherited dependency	24
Different types of inherited dependencies	8
Total	198

Figure 7: detected errors concerning paradigmatic links

8. Conclusion

The relatively high score of remaining errors after double annotation and validation suggests that well-formedness rules should play an important role in any annotation process. They can be used post-validation, as was the case here, or they can even be included in the annotation system, rendering erroneous annotations impossible. It has to be noted, however, that many rules are not as categorical as those presented here, but are rather of a statistical nature. For example, adverbs can be subjects of verbs (*Today is a good day*), but due to their sporadic nature, links of this type should be allowed and checked manually in the final stage of annotation.

High accuracy and coherence of the annotation is not just a theoretical issue. The improvement of today's statistical

parsers depends crucially on the gold-standard of the treebank on which the parser is trained. Ongoing work explores the relation between parse errors and the corresponding difficulties encountered in human annotation on the one hand and the goal to improve the annotation schemes on the other, in particular concerning the various possible tree structures for representing coordination.

9. References

De Marneffe M. C., Manning C. D. (2008). Stanford typed dependencies manual. http://nlp.stanford. edu/software/dependencies manual. Pdf.

Deulofeu J., Dufort L., Gerdes K., Kahane S., Pietrandrea P. (2010) Depends on what the French say: Spoken corpus annotation with and beyond syntactic function, *The Fourth Linguistic Annotation Workshop (LAW IV),* Uppsala, 8 p.

Gerdes K., Kahane S. (2009) Speaking in piles: Paradigmatic annotation of French spoken corpus, *Processing of the fifth Corpus Linguistics Conference*, Liverpool, 15 p.

Gerdes K., Kahane S., Lacheret A., Pietrandrea P., Truong A. (2012) Intonosyntactic data structures: the Rhapsodie treebank of spoken French, *Proceedings of the Sixth Linguistic Annotation Workshop*, pp. 85-94, Jeju.

Gerdes, K. (2013) Collaborative Dependency Annotation, *Proceedings of Depling 2013*, 88–97, Prague.

Hoekstra H., Moortgat M., Renmans B., Schouppe M., Schuurman I., Van der Wouden T. (2003), CGN Syntactische Annotatie. http://lands.let.kun.nl/cgn/doc_Dutch/topics/version_1. 0/annot/syntax/ syn_prot.pdf

Johansson, R. Dependency Syntax in the CoNLL Shared Task 2008. http://faculty.ist.unomaha.edu/ylierler/ teaching/material/conll-syntax.pdf

Kahane S. (2001) Grammaires de dépendance formelles et théorie Sens-Texte, Tutoriel, *Proceedings of TALN*, vol. 2, Tours, 63 p.

Kahane S. (2012) De l'analyse en grille à la modélisation des entassements, in S. Caddeo, M.-N. Roubaud, M. Rouquier, F. Sabio, *Penser les langues avec Claire Blanche-Benveniste*, Presses de l'université de Provence, 101-116.

Kahane S., Pietrandrea P. (2012) La typologie des entassements en français, *Actes du 3ème congrès mondial de linguistique française* (*CMLF*), Lyon, 1809-1828.

Lacheret A., Kahane S., Pietrandrea P., Avanzi M., Victorri B. (2011), Oui mais elle est où la coupure, là ? Quand syntaxe et prosodie s'entraident ou se complètent, in F. Lefeuvre & E. Moline (éd.), *Unités syntaxiques et unités prosodiques*, *Langue Française*.

Lacheret A., Beliao J., Dister A., Gerdes K., Goldman J.-P., Kahane S., Obin N. Tchobanov A. (2014) Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French, *Proceedings of LREC 2014,* Reykjavik.

Mel'čuk I. (1988) *Dependency Syntax : Theory and Practice*, SUNY Press.

Mel'čuk, I., & Pertsov, N. (1987). *Surface syntax of English. A Formal Model within the Meaning-Text Framework*, Amsterdam: Benjamins.

Nivre J., Hall J., Kübler S. McDonald R. Nilsson J., Riedel S., Deniz Yuret D. (2007). The CoNLL 2007 shared task on dependency parsing. *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*.

Tesnière L. (1959) *Eléments de syntaxe structurale*, Klincksieck.

Villemonte de la Clergerie E. (2010) *Building factorized TAGs with meta-grammars, Proceedings of the 10th International Conference on Tree Adjoining Grammars and Related Formalisms - TAG+10*, 111-118.