Description of the Rhapsodie TreeBank's Tabular Format

Version: morpho-syntax, micro-syntax

June 28, 2015

Document authors: Rachel Bawden and Ilaine Wang
Creation of the tabular format: Rachel Bawden, Ilaine Wang, with the collaboration of Julie Belião
Coordination: Kim Gerdes, Sylvain Kahane
Annotation Platform (Arborator): Kim Gerdes
Micro-syntactic annotation: Rachel Bawden, Christophe Benzitoun, Marie-Amélie Botalla, Adèle Désoyer, Sylvain Kahane, Paola Pietrandrea
Macro-syntactic annotation: Christophe Benzitoun, Jeanne-Marie Debaisieux, José Deulofeu, Anne Dister, Florence Lefeuvre, Paola Pietrandrea, Nathalie Rossi-Gensane, Frédéric Sabio, Noalig Tanguy, Bernard Victorri
Prosody: Mathieu Avanzi, Julie Belião, Jean-Philippe Goldman, Anne Lacheret-Dujour, Philippe Martin, Nicolas Obin, Arthur Truong, Bernard Victorri

---

Certain samples used in the Rhapsodie project have been taken from preexisting external data but are identified here under a generic name (Rhap_sample_number). The association of these samples to their primary sources can be consulted at http://projet-rhapsodie.fr/propriete-intellectuelle.html.

In addition to the tabular file for each text (Rhap-D0001.micro.tabular, Rhap-M2006.micro.tabular etc.), there is a global tabular file (Rhapsodie.micro.tabular), which contains the information from all of the texts.

Columns 1-5 contain technical information.
Columns 6-14 contain morpho-syntactic information.
Columns 15-27 contain micro-syntactic information.

Technical Columns

1. *Text_ID*: the text name (D0001, M2006 etc.)
2. *Tree_ID*: the number of the tree in the text
3. *Token_ID*: the number of the token in the tree
4. *Token*: the form of the token. Lexemes made up of several orthographic words have been segmented into individual tokens. A token is therefore a segment of the transcription found between two whitespaces or a whitespace and a punctuation

mark. All non-alphabetic characters (spaces, dashes and apostrophes) are also considered individual tokens.

5. *Speaker*: The speaker ID. Where overlapping occurs, there can be several speakers (annotated for example as `$L1-$L3`).

Morpho-syntactic Columns

6. *Word_span*: the position of the token in the wordform. The value is either `B` (begin) for the first token of the word or `I` (inner) for all tokens that are not the first token of the word.

7. *Wordform*: the wordform to which the token belongs. When a wordform is made up of several tokens, this is only indicated for the first token.

8. *Lemma*: the lemma of the wordform to which the token belongs. When there are several tokens that belong to the same wordform, the lemma is not repeated; the lemma is written in the row corresponding to the first token.

9. *POS*: the morpho-syntactic category associated with the wordform to which the token belongs. The possible values are `N`, `V`, `Adj`, `Adv`, `I`, `Pre`, `D`, `Cl`, `Pro`, `CS`, `Qu`, `J`, `Pre+D`, `Pre+Qu` or `X` (for unknown categories).

10. *Mood*: verbal mood, with 5 possible values: `indicative`, `subjunctive`, `imperative`, `infinitive`, `past_participle` and `present_participle`. When the form is ambiguous, the two modal possibilities are indicated (e.g. `indicative/subjunctive`)

11. *Tense*: verbal tense with 5 possible values: `present`, `future`, `conditional`, `imperfect` and `perfect`. Tense is only marked for verbs whose mood is `indicative`.

12. *Person*: grammatical person for verbs and personal pronouns, with 3 possible values: `1`, `2` or `3`. When the form is ambiguous, all the possible values are noted, separated by slashes (e.g. `1/2/3`).

13. *Number*: grammatical number (`sg` or `pl` or `sg/pl` when there is ambiguity) for inflected verbs, nouns, adjectives and certain *qu-* words (quel, quels, laquelle etc.).

14. *Gender*: grammatical gender (`masc`, `fem` or `masc/fem` when there is ambiguity) for nouns, adjectives, past participles and certain *qu-* words (quel, quels, laquelle etc.).

Micro-syntactic Columns

The two columns 15 and 16 contain exactly one dependency link for each wordform. They contain an independent and complete dependency analysis based on the dependency links in the following columns (17 to 27).

15. *ID_dep*: the number of the governor by dependency. The number of the governor corresponds to the column Token_ID. When the dependent is made up of several tokens, the relation is only written for the first token. When the governor is made up of several tokens, this corresponds to the Token_ID of the first token of the governor.

    This principle also holds for the other micro-syntactic columns.

16. *Type_dep*: the type of dependency link corresponding to ID dep.

The columns 17-27 correspond to the individual classes of dependency link ('plain', 'para', 'inherited', 'junc', 'junc_inherited'). The first column of each pair corresponds to the governor number and the second to the type of link.

17. *ID_plain*: the number of the governor by 'plain' dependency.

18. *Type_plain*: the type of the (plain) dependency relation corresponding to ID_plain (with the possible functions `pred`, `root`, `sub`, `dep`, `obj`, `obl`, `ad`).

    N.B. There can only be a single type of plain dependency and a single plain governor per token.

19. *ID_junc*: the number of the governor by the `junc` relation (by junction).

20. *Type_junc*: the type of junction relation - there is only one, so the only possible value is `junc`. This column is present for the uniformity of the table.

21. *ID_para*: the number of the governor by paradigmatic relation.

22. *Type_para*: the type of paradigmatic link (out of `para_disfl`, `para_coord`, `para_intens`, `para_dform`, `para_reform`, `para_hyper`, `para_negot`).

    N.B. there can only be a single type of paradigmatic dependency and a single paradigmatic governor per token.

23. *ID_inherited*: the number of the governor by inherited dependency.

24. *Type_inherited*: the type of inherited dependency (out of `pred_inherited`, `root_inherited`, `sub_inherited`, `dep_inherited`, `obj_inherited`, `obl_inherited`, `ad_inherited`).

    N.B. there can only be a single type of dependency per token, but a token can have several governors by inherited dependency. In this case, the numbers of the governors are separated by a comma.

E.g.

Token_ID	Token	ID_para	Type_para	ID_inher	Type_inher
5	de	3			
6					
7	de	5	para_disfl	3	obl_inherited
8					
9	de	7	para_disfl	3	obl_inherited
10					
11	quotidien	9		5.7	dep_inherited

25. *ID_junc_inherited*: the number of the governor by inherited junction.

26. *Type_junc_inherited*: the type of inherited junction - there is only one possible value (`junc_inherited`). This column is present for the uniformity of the table.

27. *Layer*: indicates that the token belongs to a layer. In this annotation, different levels of layer are flattened. The example '{ c'est un | c'est une { des | des } | c'est une des } mesures du plan banlieue' would be represented as follows:

Text_ID	Token_ID	Token	Layer
D0002	21	c	B
D0002	22	'	I
D0002	23	est	I
D0002	24		
D0002	25	une	I
D0002	26		
D0002	27	des	U
D0002	28		
D0002	29	&	
D0002	30		
D0002	31	des	U
D0002	32		
D0002	33	&	
D0002	34		
D0002	35	c	B
D0002	36	'	I
D0002	37	est	I
D0002	38		
D0002	39	une	I
D0002	40		
D0002	41	des	L
D0002	42		
D0002	43	mesures	O

Additional comments:

The contractions 'au', 'aux', 'du', 'des', 'auquel', 'auxquels' etc. of the form Pre + D are not segmented into two tokens 'à + le', 'à + les' etc. in the tabular format. However the lemma contains the two forms and the part of speech contains the two morpho-syntactic categories.

E.g.

Text_ID	Tree_ID	Token_ID	Token	Wordform	Lemma	POS
D2011	94	11	des	de+les	de+le	Pre+D
D2011	94	12				
D2011	94	13	odeurs	odeurs	odeur	N

Detailed coding guides for micro-syntactic and macro-syntactic annotations are available on the project's tutorial page: http://projet-rhapsodie.fr/plus/tutoriels.html.