Description of the Rhapsodie TreeBank's Tabular Format

Version: morpho-syntax, micro-syntax, macro-syntax, prosody

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_macro_prosody.tabular, Rhap-M2006.micro_macro_prosody.tabular etc.), there is a global tabular file (Rhapsodie.micro_macro_prosody.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.
Columns 28-40 contain macro-syntactic information.
Columns 41-63 contain prosodic 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

Macro-syntactic Columns

All macro-syntactic information is given in BILOU format (corresponding to Begin, In, Last, Out, Unique). The given value indicates the position of the token in the IU, if it belongs to an IU (`B` at the beginning, `I` in the middle, `L` at the end and `U` when the token is an IU in its own right) and if it does not belong to an IU, it has the value `O`.

28. *IU*: the token's membership to an illocutionary unit. This information is provided in the BILOU format, however it should be noted that the value `O` is not used here since all tokens must necessarily belong to an IU.

29. *Nucleus*: the token's membership to a nucleus.

30. *Prenucleus*: the token's membership to a pre-nucleus.

31. *Gov_prenucleus*: the token's membership to a governed pre-nucleus, i.e. its membership to the same government unit as the IU to which it is attached.

32. *Innucleus*: the token's membership to an in-nucleus.

33. *Gov_innucleus*: the token's membership to a governed in-nucleus.

34. *Postnucleus*: the token's membership to a post-nucleus.

35. *Gov_postnucleus*: the token's membership to a governed post-nucleus.

36. *IU_parenthesis*: the token's membership to a parenthetic IU.

37. *IU_graft*: the token's membership to a graft, which generally corresponds to reported speech or to sequences which contribute to filling the syntactic position of a governing element.

38. *IU_embedded*: the token's membership to an embedded unit which does not correspond to an IU (whilst it is the case of a graft).

39. *Associative_nucleus*: the token's membership to a closed list of discursive markers drawn up for the project.

40. *Intro_IU*: the token's membership to an IU initiator.

Prosodic Columns

Columns 41 to 56 are secondary data (from annotations), whereas columns 57 to 63 are primary data (extracted directly from the signal).

The period, package, rhythmic group and foot are different units used in the project's prosodic annotations. For each unit, we have included an additional attribute describing the tonal profile (unit_*tone*). The value of this attribute represents the F0 height according to a speaker average and can be one of five levels: `L` for very low, `l` for low, `m` for middle, `h` for high, et `H` for very high. This description is given for the start and end of the unit. To these values are added the level and position of the most prominent point in the unit, which can be in the first third (`1`), the second (`2`) ou or in the third (`3`).

E.g. The value `lhH3` assigned to a syllable means that the start of the F0 contour is low, it ends high and that the most prominent point is very high and is associated with the final third of the syllable.

41. *Period*: the token's membership to an intonative period. The values are given in BILOU format (described in the introduction to the preceding section), without `O` as was the case for the IU (28), since each token must belong to an intonative period. Sometimes an additional value is found: as well as marking the position of the token in the period, each value can also describe whether the period is truncated to the left (`-B`, `-I`, `-L`, `-U`), to the right (`B-`,`I-`,`L-`,`U-`), or on both sides (`-B-`,`-I-`,`-L-`,`-U-`).

42. *Period_tone*: the tonal profile (or intonation contour) of the intonative period to which the token belongs.

43. *Package*: delimitation of intonative packages into units in BILOU format, as for periods and also without the value `O` since every token must belong to an intonative package.

44. *Package_type*: the type of intonative package to which the token belongs out of the following values: `filled-dis`, `filled-pause`, `included`, `lone`, `lone-dis-strong`, `motherless`, `motherless-dis-weak`, `silent-pause`, or `tail`.

45. *Package_tone*: the tonal profile (or intonation contour) of the intonative package to which the token belongs.

46. *Group*: delimitation of rhythmic groupes into units in BILOU format (without the value `O` since every token must belong to a rhythmic group).

47. *Group_type*: the type of rhythmic group to which the token belongs out of the following values: `dis-strong`, `dis-weak`, `filled-dis`, `filled-pause`, `silent-pause`, `strong`, `tail`, or `weak`.

48. *Group_tone*: the tonal profile (or intonation contour) of the rhythmic group to

which the token belongs.

49. *Foot*: delimitation into metrical foot units in BILOU format (without the value `O`)

50. *Foot_type*: the last metrical foot of the token. As with the rhythmic group, it can be annotated `dis-strong`, `dis-weak`, `filled-dis`, `filled-pause`, `silent-pause`, `strong`, `tail`, or `weak`.

51. *Foot_tone*: the tonal profile (or intonative contour) of the last metrical foot of the token.

52. *Syllable*: delimitation into the syllabic units in BILOU format analysed in Syllable_tone. The value `O` again does not apply here. Given that here we only study the last syllable of each token, we note just U for a token made up of a word or a syllable, but we regroup several words that make a single token (d+'+abord), and several tokens making up a single syllable (e.g. de la [dla]).

53. *Syllable_tone*: the tonal profile (or intonative contour) of the last syllable of the token.

54. *Prominence_initial*: the degree of prominence of the first syllable of the token. A prominence can be annotated `W` for *Weak* or `S` for *Strong*. Where the syllable is not prominent, it can be annotated `O`, `_` (pause) or `%` (for an inaudible syllable or one that has not been transcribed because of overlapping).

55. *Prominence_final*: the degree of prominence of the final syllable of the token. The values are the same as for initial prominence.

56. *Hesitation*: annotated `H` for the particle 'euh' or for a hesitant syllable, `_` (pause) or `%` (for an inaudible syllable or one that has not been transcribed because of overlapping).

57. *Pause_length*: the length of the pause (in seconds), annotated for the preceding token. In other words, if the value is empty for a given token, then the token is not followed by a pause. It should be noted that in the case of an overlap and if the main speaker's speech is transcribed first, the length of the pause is indicated on the last word at the end of the overlap, but the symbol `#` (hash) can be found in the row of the token after which the pause was really observed. This format was chosen in the interest of ensuring a pertinent alignment with the different units of analysis, notably the period.

58. *Tmin*: the time code of the beginning of the token (or where appropriate, the overlap) within the text.

59. *Tmax*: the time code of the end of each token (or where appropriate, the overlap) within the text.

60. *Syllable_length*: the length of the last syllable of the token (in milliseconds).

61. *Syllable_length_avg*: the average length of the syllable (in milliseconds).

62. *Pitch*: the height (in semi-tones) of the syllable.

63. *Pitch_avg*: the average height (in semi-tones), calculated over a span of several syllables preceding and following the last syllable of the token.

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.