Description du format tabulaire du TreeBank Rhapsodie

Version : morpho-syntaxe, micro-syntaxe

28 avril 2015

Rédaction du document : Rachel Bawden et Ilaine Wang
Création des fichiers tabulaires : Rachel Bawden, Ilaine Wang, avec la collaboration de Julie Belião
Coordination : Kim Gerdes, Sylvain Kahane
Plateforme d'annotation (Arborator) : Kim Gerdes
Annotation microsyntaxique : Rachel Bawden, Christophe Benzitoun, Marie-Amélie Botalla, Adèle Désoyer, Sylvain Kahane, Paola Pietrandrea
Annotation macrosyntaxique : Christophe Benzitoun, Jeanne-Marie Debaisieux, José Deulofeu, Anne Dister, Florence Lefeuvre, Paola Pietrandrea, Nathalie Rossi-Gensane, Frédéric Sabio, Noalig Tanguy, Bernard Victorri
Prosodie : Mathieu Avanzi, Julie Belião, Jean-Philippe Goldman, Anne Lacheret-Dujour, Philippe Martin, Nicolas Obin, Arthur Truong, Bernard Victorri

---

Certains échantillons utilisés dans le projet Rhapsodie sont issus de données préexistantes externes mais sont identifiés ici sous un nom générique (Rhap_numéro d'échantillon). Les correspondances avec les sources primaires peuvent être consultées sur cette page http://projet-rhapsodie.fr/propriete-intellectuelle.html.

Il existe un fichier tabulaire global (Rhapsodie.micro.tabular), contenant tous les textes ainsi qu'un fichier tabulaire pour chaque texte (ex : Rhap-D0001.micro.tabular, Rhap-M2006.micro.tabular etc.).

Les colonnes 1-5 correspondent aux informations techniques.
Les colonnes 6-14 correspondent aux informations morpho-syntaxiques.
Les colonnes 15-27 correspondent aux informations micro-syntaxiques.

Colonnes techniques

1. *Text_ID* : le nom du texte (D0001, M2006 etc.)
2. *Tree_ID* : le numéro de l'arbre dans le texte
3. *Token_ID* : le numéro du token dans l'arbre
4. *Token* : la forme du token. Des lexèmes composés de plusieurs mots orthographiques ont été segmentés en tokens individuels. Un token est donc un segment

de la transcription compris entre deux blancs ou un blanc et un signe de ponctuation. Tous les caractères qui ne sont pas des lettres (les espaces, les tirets et les apostrophes) sont considérés comme des tokens individuels aussi.

5. *Speaker* : l'identifiant du locuteur. En cas de chevauchement, on peut avoir plusieurs locuteurs (annotés alors `\$L1-\$L3` par exemple).

Colonnes morpho-syntaxiques

6. *Word_span* : la position du token dans le mot forme. La valeur est soit `B` (begin) pour le premier token d'un mot, soit `I` (inner) pour les tokens qui en sont pas les premiers tokens du mot.
7. *Wordform* : le mot-forme auquel appartient le token. Dans le cas d'un mot-forme comprenant plusieurs tokens, le mot-forme est uniquement marqué pour le premier token.
8. *Lemma* : le lemme du lexème auquel appartient le token. Dans le cas où il y a plusieurs tokens qui font partie du même lexème, le lemme n'est pas répété : le lemme est écrit dans cette colonne pour la ligne correspondant au premier token du lexème.
9. *POS* : la catégorie morpho-syntaxique du mot auquel appartient le token parmi `N`, `V`, `Adj`, `Adv`, `I`, `Pre`, `D`, `Cl`, `Pro`, `CS`, `Qu`, `J`, `Pre+D`, `Pre+Qu` ou `X` (pour les catégories inconnues).
10. *Mood* : le mode pour les verbes parmi `indicative`, `subjunctive`, `imperative`, `infinitive`, `past_participle` et `present_participle`. Dans le cas où la forme est ambiguë, les deux possibilités de mode sont indiquées (ex : `indicative/subjunctive`).
11. *Tense* : le temps grammatical du verbe parmi `present`, `future`, `conditional`, `imperfect` et `perfect`. Le temps est marqué uniquement pour les verbes qui ont pour mode `indicative`.
12. *Person* : la personne grammaticale pour les verbes et les pronoms personnels, (`1`, `2` ou `3`). En cas d'ambiguïté, les personnes possibles sont toutes écrites séparées par des barres obliques (ex : `1/2/3`).
13. *Number* : le nombre grammatical (`sg` ou `pl` ou `sg/pl` en cas d'ambiguïté) pour les verbes conjugués, les noms, les adjectifs, les pronoms et certains mots *qu-* (quel, quels, laquelle etc.).
14. *Gender* : le genre grammatical (`masc`, `fem` ou `masc/fem` en cas d'ambiguïté) pour les noms, les adjectifs, les participes passés et certains mots *qu-* (quel, quels, laquelle etc.).

Colonnes micro-syntaxiques

Les deux colonnes 15 et 16 contiennent exactement un lien pour chaque mot-forme. Il s'agit d'une colonne indépendante qui contient une analyse en dépendance complète.

15. *ID_dep* : le numéro du gouverneur par dépendance. Le numéro du gouverneur correspond à la colonne Token_ID. Dans le cas où un gouverneur est constitué de plusieurs tokens, c'est le Token_ID du premier token qui est pris comme numéro de gouverneur. Ce principe tient aussi pour les autres types de liens de dépendance.
16. *Type_dep* : le type de lien de dépendance correspondant à ID dep.

Les colonnes 17-27 correspondent aux classes individuelles de lien de dépendance ("plain", "para", "inherited", "junc", "junc_inherited"). La première colonne de chaque paire correspond au numéro du gouverneur et la seconde au type de lien.

17. *ID_plain* : le numéro du gouverneur par dépendance "primitive".
18. *Type_plain* : le type de lien de dépendance (primitif), correspondant à ID_plain (correspondant aux fonctions `pred`, `root`, `sub`, `dep`, `obj`, `obl`, `ad`).
    N.B. Il ne peut y avoir qu'un seul type de dépendance primitive et un seul gouverneur primitif par token.
19. *ID_junc* : le numéro du gouverneur par lien "junc" (de jonction)
20. *Type_junc* : le type de lien junc - il n'y en a qu'un seul, donc ceci correspond toujours à junc. Cette colonne est ici pour l'uniformité du tableau.
21. *ID_para* : le numéro du gouverneur par lien paradigmatique.
22. *Type_para* : le type de lien paradigmatique (parmi les types `para_disfl`, `para_coord`, `para_intens`, `para_dform`, `para_reform`, `para_hyper`, `para_negot`)
    N.B. Il ne peut y avoir qu'un seul type de dépendance paradigmatique et un seul gouverneur paradigmatique par token.
23. *ID_inherited* : le numéro du gouverneur par lien hérité.
24. *Type_inherited* : le type de lien hérité (parmi `pred_inherited`, `root_inherited`, `sub_inherited`, `dep_inherited`, `obj_inherited`, `obl_inherited`, `ad_inherited`).
    N.B. Il ne peut y avoir qu'un seul type de dépendance par token, mais il peut y avoir plusieurs gouverneurs par dépendance héritée. Dans ce cas, les numéros des gouverneurs sont séparés par une virgule. Ex :

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* : le numéro du gouverneur par lien “junc_inherited” (de jonction héritée)
26. *Type_junc_inherited* : le type de lien junc_inherited - il n’y en a qu’un seul, donc ceci correspond toujours à `junc_inherited`. Cette colonne est ici pour l’uniformité du tableau.
27. *Layer* : l’appartenance à un entassement. Dans cette annotation, les différents niveaux d’entassement sont écrasés. On aura donc pour l’exemple “{ c’est un | c’est une { des | des } | c’est une des } mesures du plan banlieue” l’annotation suivante :

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

Remarques supplémentaires :

Les amalgames “au”, “aux”, “du”, “des’, “auquel”, “auxquels” etc. en Pre + D ne sont pas segmentés en deux tokens “à + le”, “à + les” etc. dans le format tabulaire. Par contre, le lemme indique les deux formes, et la catégorie morpho-syntaxique contient les deux catégories morpho-syntaxiques.

Ex :

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

Des protocoles de codage détaillés pour les annotations micro-syntaxiques et macro-syntaxiques sont disponibles sur la page des tutoriels du projet : http://projet-rhapsodie.fr/plus/tutoriels.html.