Description du format tabulaire du TreeBank Rhapsodie

Version : morpho-syntaxe, micro-syntaxe, macro-syntaxe, prosodie

2 juillet 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_macro_prosody.tabular), contenant tous les textes ainsi qu'un fichier tabulaire pour chaque texte (ex : Rhap-D0001.micro_macro_prosody.tabular, Rhap-M2006.micro_macro_prosody.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.
Les colonnes 28-40 correspondent aux informations macro-syntaxiques.
Les colonnes 41-63 correspondent aux informations prosodiques.

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 orthogra-

phiques 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

Colonnes macro-syntaxiques

Toutes les informations macro-syntaxiques sont données au format BILOU (correspondant à Begin, In, Last, Out, Unique). La valeur associée à chaque token indique sa position au sein de chacune des unités macro-syntaxiques : il sera ainsi annoté `B` s'il en est le début, `I` s'il ne se trouve à aucune des extrêmités de l'unité, `L` s'il est à la fin, ou encore `U` dans le cas où le token constitue à lui seul une unité). En revanche, si le token donné ne fait pas partie de l'unité macro-syntaxique en question, on lui associe alors la valeur `O`.

28. *IU* : l'appartenance à une unité illocutoire (UI, ou *Illocutionary Unit* en anglais). Cette information est bien fournie au format BILOU, mais on note que pour cette unité, la valeur `O` n'est pas utilisée du fait que tous les tokens font nécessairement partie d'une UI.
29. *Nucleus* : l'appartenance à un noyau.
30. *Prenucleus* : l'appartenance à un pré-noyau.
31. *Gov_prenucleus* : l'appartenance à un pré-noyau régi, c'est-à-dire qui fait partie de la même unité rectionnelle que l'UI à laquelle il est rattaché.
32. *Innucleus* : l'appartenance à un in-noyau.
33. *Gov_innucleus* : l'appartenance à un in-noyau régi.
34. *Postnucleus* : l'appartenance à un post-noyau.
35. *Gov_postnucleus* : l'appartenance à un post-noyau régi.
36. *IU_parenthesis* : l'appartenance à une UI parenthétique.
37. *IU_graft* : l'appartenance à une greffe, qui correspond généralement à un discours rapporté ou bien à des séquences qui contribuent à remplir la position syntaxique d'un élément recteur.
38. *IU_embedded* : l'appartenance à une unité enchâssée qui ne correspond pas à une UI (alors que c'est le cas de la greffe).
39. *Associative_nucleus* : l'appartenance à une liste fermée de marqueurs discursifs établie pour le projet.
40. *Intro_IU* : l'appartenance à une unité qui permet d'introduire une UI.

Colonnes prosodiques

Les colonnes 41 à 56 sont des données secondaires (issues d'annotations), tandis que les colonnes 57 à 63 sont des données primaires (extraites directement à partir du signal).

La période, le paquet, le groupe rythmique et le pied sont les différentes unités sur lesquelles reposent les annotations prosodiques du projet. Pour chacune d'elle, nous avons donc un attribut supplémentaire décrivant leur profil tonal (unité_*tone*). Ce dernier a pour valeur la description du niveau de la hauteur de la F0 par rapport à la moyenne du locuteur selon un encodage sur cinq niveaux : `L` pour un niveau très bas, `l` pour un niveau bas, `m` pour un niveau moyen, `h` pour un niveau haut, et `H` pour un niveau très haut. Cette description est donnée pour le point initial de l'unité, et son point final. À cette information peut s'ajouter le niveau et la position du point le plus saillant (extremum) dans l'unité donnée, qui peut être dans le premier tiers `1`, dans le deuxième `2` ou dans le dernier `3`.

Ex : La valeur `lhH3` sur une syllabe signifie donc que le point initial du contour de F0 de la syllabe est bas, que son point final est haut, et que le point le plus saillant est très haut et se trouve dans le dernier tiers de la syllabe.

41. *Period* : l'appartenance à une période intonative. Cette information est donnée au format BILOU (décrit dans l'introduction de la section précédente) sans le `O` comme c'était le cas pour l'UI (28), puisque, de la même manière, tout token fait partie d'une période intonative. On y trouve toutefois un détail supplémentaire : en plus de marquer la position du token dans la période, chaque valeur peut également indiquer s'il s'agit d'une période tronquée à gauche (`-B`, `-I`, `-L`, `-U`), à droite (`B-`,`I-`,`L-`,`U-`), des deux côtés (`-B-`,`-I-`,`-L-`,`-U-`).
42. *Period_tone* : le profil tonal (ou contour intonatif) de la période intonative dans laquelle se trouve le token.
43. *Package* : délimitation des paquets intonatifs en unités au format BILOU (sur le schéma des périodes et toujours sans le `O` puisque tout token fait forcément partie d'un paquet intonatif).
44. *Package_type* : le type de paquet intonatif dans lequel le token se trouve parmi `filled-dis`, `filled-pause`, `included`, `lone`, `lone-dis-strong`, `motherless`, `motherless-dis-weak`, `silent-pause`, ou `tail`.
45. *Package_tone* : le profil tonal (ou contour intonatif) du paquet intonatif dans lequel se trouve le token.
46. *Group* : délimitation des groupes rythmiques en unités au format BILOU (toujours sans le `O` puisque tout token fait forcément partie d'un groupe rythmique).
47. *Group_type* : le type de groupe rythmique dans lequel le token se trouve parmi `dis-strong`, `dis-weak`, `filled-dis`, `filled-pause`, `silent-pause`, `strong`, `tail`, ou `weak`.
48. *Group_tone* : le profil tonal (ou contour intonatif) du groupe rythmique dans lequel se trouve le token.

49. *Foot* : délimitation des unités de pieds métriques au format BILOU (toujours sans le `O`).
50. *Foot_type* : le type du dernier pied métrique du token. À l'instar du groupe rythmique, il peut être annoté `dis-strong`, `dis-weak`, `filled-dis`, `filled-pause`, `silent-pause`, `strong`, `tail`, ou `weak`.
51. *Foot_tone* : le profil tonal (ou contour intonatif) du dernier pied métrique du token.
52. *Syllable* : délimitation des unités syllabiques au format BILOU (toujours sans le `O`) analysées dans Syllable_tone. Étant donné que l'on n'étudie dans ce tabulaire que la dernière syllabe de chaque token, on note simplement U pour un token constitué d'un mot ou d'une syllabe, mais on regroupe plusieurs mots constituant un seul token (d+'+abord), ou encore plusieurs tokens constituant une unique syllabe (ex : de la [dla]).
53. *Syllable_tone* : le profil tonal (ou contour intonatif) de la dernière syllabe du token.
54. *Prominence_initial* : le degré de proéminence de la première syllabe du token. Une proéminence peut être annotée `W` pour *Weak* ou `S` pour *Strong*. Dans le cas où cette syllabe n'est pas proéminente, elle peut avoir comme valeur `0`, `_` (pause) ou encore `%` (syllabe inaudible ou non transcrite en raison d'un chevauchement).
55. *Prominence_final* : le degré de proéminence de la dernière syllabe du token. Elle peut avoir les mêmes valeurs que pour la proéminence initiale.
56. *Hesitation* : marquée `H` pour la particule "euh" ou bien pour une syllabe hésitante, `_` (pause) ou bien `%` (mot inaudible ou non transcrit en raison d'un chevauchement).
57. *Pause_length* : la durée de la pause (en s), indiquée au niveau du token qui la précède. Autrement dit, si cette donnée est vide pour un token donné, alors il n'est pas suivi d'une pause. À noter que dans le cas précis d'un chevauchement et si les paroles du locuteur principal sont transcrites en premier, la longueur de la pause est indiquée sur le dernier mot, à la fin du chevauchement mais on a ajouté le signe `#` (dièse) sur la ligne du mot après lequel on observe réellement une pause. Ce choix a été fait afin de préserver un alignement pertinent avec les différentes unités d'analyse, notamment la période.
58. *Tmin* : le temps de début de chaque token (ou le cas échéant, du chevauchement) au sein de son échantillon.
59. *Tmax* : le temps de fin de chaque token (ou le cas échéant, du chevauchement) au sein de son échantillon.
60. *Syllable_length* : la durée de la dernière syllabe du token (en ms).
61. *Syllable_length_avg* : la durée moyenne de cette syllabe (en ms).
62. *Pitch* : la hauteur (en demi-tons) de cette syllabe.
63. *Pitch_avg* : la hauteur moyenne (en demi-tons), calculée sur un empan de quelques syllabes précédant et suivant la dernière syllabe du token.

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.