Annotations *Rhapsodie*

pour le *Trameur* (v3)

06/07/2014 19:45:10

Serge Fleury

(document de travail)

Références

Le Trameur, manuel d'utilisation

http://www.tal.univ-paris3.fr/trameur/leMetierLexicometrique.pdf

Dans cette documentation, la partie « *Relations de dépendance entre les items de Trame (via leurs annotations)* » présente l'état des développements actuellement disponibles dans le *Trameur* pour l'exploitation des relations de dépendance.

Le Trameur. Propositions de description et d'implémentation des objets textométriques

http://www.tal.univ-paris3.fr/trameur/trameur-propositions-definitions-objets-textometriques.pdf

Ce document met au jour une description des objets textométriques et les méthodes mises en œuvre dans le ***Trameur*** pour travailler sur et avec ces objets dans une perspective textométrique. On y détaille aussi les opérations permises sur une *base textométrique* : format des données textuelles, modification dynamique de la *Trame*, correction ou ajout d'annotation etc.

Annotations Rhapsodie pour le Trameur

http://www.tal.univ-paris3.fr/trameur/bases/rhapsodie2trameur.pdf

Présentation du processus de transcodage des annotations Rhapsodie pour construire une base Textométrique ; présentation des différents processus de traitements des annotations de dépendance.

1. Préambule

Ce document présente le processus de transcodage des données issues du projet ***Rhapsodie*** (http://projet-rhapsodie.fr/) sous la forme d'une base textométrique importable dans le *Trameur*.

2. Données Rhapsodie

Le fichier d'annotations (`Rhapsodie.micro_macro_prosody`) est visible dans la figure ci-contre (lecture ici dans un tableur).

Le fichier tabulé regroupant les informations Rhapsodie est composé de 48 colonnes décrites infra.

Text ID	Tree ID	Token	Token	Words	Word fe	Lemma	POS	Mood	Tense	Person	Number	Gender	ID dep	Type de	ID plair	Type pl	ID junc	Type ju	ID para	Type pa	ID inhe	Type in	ID junc	Type ju	Layer	IU	Nucleu	Prenuc	Gov nu	Innucle	Gov inn	Postnu	Gov poi	IU pare	IU graft	IU emb	Associe	Intro IU	Fromin	Fromin	Hesitat	Pitch a	Pitch	Syllabl	Syllabl	Speake	Pause length
M2004																																															
M2004	1	1	mes	B	mes	son	D				pl	masc	5	dep	5	dep									0	B	0	B	0	0	0	0	0	0	0	0	0	0	0		88.07	85.29	85	145	S1	
M2004	1	2																																													
M2004	1	3	chers	B	chers	cher	Adj				pl	masc	5	dep	5	dep									0	I	0	I	0	0	0	0	0	0	0	0	0	S	S		84.85	89.35	200	91.67	S1	
M2004	1	4																																												
M2004	1	5	compat	B	compat	compat	N				pl	masc	0	root	0	root									0	I	0	L	0	0	0	0	0	0	0	0	0	S		0	80.42	74.36	393.3	150.2	S1	0.872
M2004	1	6																																												
M2004	1	7	je	B	je	je	Cl			1	sg		9	sub	9	sub									0	I	B	0	0	0	0	0	0	0	0	0	0	_		84.99	80.92	75	100	S1	
M2004	1	8																																												
M2004	1	9	voudra	B	voudra	vouloir	V	indicat	conditi	1	sg		0	root	0	root									0	I	I	0	0	0	0	0	0	0	0	0	W		0	82.41	86.59	110	203.7	S1	
M2004	1	10																																												
M2004	1	11	d	B	d'abord	d'abord	Adv						0	root	0	root									0	I	I	0	0	0	0	0	0	0	0	0	S		0	85.54	80.86	560	100	S1	
M2004	1	12	'		I																				0	I	I	0	0	0	0	0	0	0	0	0	S		0	85.54	80.86	560	100	S1	
M2004	1	13	abord	I																					0	I	I	0	0	0	0	0	0	0	0	0	S		0	85.54	80.86	560	100	S1	
M2004	1	14																																												
M2004	1	15	exprim	B	exprim	exprim	V	infinitive					9	obj	9	obj									0	I	I	0	0	0	0	0	0	0	0	0	W	_		81.19	80.77	120	198.7	S1	
M2004	1	16																																												
M2004	1	17	ma	B	ma	son	D				sg	fem	19	dep	19	dep									0	I	I	0	0	0	0	0	0	0	0	0	0		0	79.57	79.53	135.4	191.1	S1	
M2004	1	18																																												
M2004	1	19	sympat	B	sympat	sympat	N				sg	fem	15	obj	15	obj									0	I	I	0	0	0	0	0	0	0	0	0	S	S		77.49	73.46	274	182.3	S1	0.819
M2004	1	20																																												
M2004	1	21	Å	B	Å	Å	Pre						15	obl	15	obl									B	I	I	0	0	0	0	0	0	0	0	0	0	_		78.72	80.49	100	387.9	S1	
M2004	1	22																																												
M2004	1	23	toutes	B	toutes	tout	D				pl	fem	25	dep	25	dep									I	I	I	0	0	0	0	0	0	0	0	0	W	W		78.66	80.6	210	332.9	S1	
M2004	1	24																																												
M2004	1	25	celles	B	celles	celui	Pro				3	pl	fem	21	dep	21	dep								L	I	I	0	0	0	0	0	0	0	0	0	S	S		80.55	76.84	565.8	155	S1	0.46
M2004	1	26																																												
M2004	1	27	et	B	et	et	J						21	junc			21	junc							B	I	0	0	0	0	0	0	0	0	0	0	0	_		80.96	79.39	112.5	121.3	S1	
M2004	1	28																																												
M2004	1	29	Å	B	Å	Å	Pre						27	junc					21	para_cl	15	obl_inherited			I	I	I	0	0	0	0	0	0	0	0	0	0	0		81.54	79.6	72.54	187.5	S1	
M2004	1	30																																												
M2004	1	31	tous	B	tous	tout	D				pl	masc	33	dep	33	dep									I	I	I	0	0	0	0	0	0	0	0	0	0	0		80.44	82.32	170	168.7	S1	
M2004	1	32																																												
M2004	1	33	ceux	B	ceux	celui	Pro				3	pl	fem	29	dep	29	dep								L	I	I	0	0	0	0	0	0	0	0	0	S	S		80.58	82.91	280	228.1	S1	
M2004	1	34																																												
M2004	1	35	qui	B	qui	qui	Ou						37	sub	37	sub									0	I	I	0	0	0	0	0	0	0	0	0	0	0		80.39	79.87	210	307.5	S1	
M2004	1	36																																												
M2004	1	37	vivent	B	vivent	vivre	V	indicat	presen	3	pl		33	dep	33	dep					25	dep_inherited			0	I	I	0	0	0	0	0	0	0	0	0	0	S	S		79.62	80.55	460	275	S1	
M2004	1	38																																												
M2004	1	39	ces	B	ces	ce	D				pl	masc	43	dep	43	dep									0	I	I	0	0	0	0	0	0	0	0	0	0	0		80.02	75.78	320	295	S1	
M2004	1	40																																												
M2004	1	41	dernier	B	dernier	dernier	Adj				pl	masc	43	dep	43	dep									0	I	I	0	0	0	0	0	0	0	0	0	0	S		77.2	79.74	220	285	S1	
M2004	1	42																																												
M2004	1	43	jours	B	jours	jour	N				pl	masc	37	obj	37	obj									0	I	I	0	0	0	0	0	0	0	0	0	0	S	S		77.56	79.01	410	197.5	S1	

1	TextID	Identifiant de l'échantillon (nom de la PARTIE dans la terminologie textométrique)
2	TreeID	Numéro de l'unité illocutoire (UI) dans l'échantillon
3	TokenID	Identifiant du *token* dans l'UI Les UI d'un échantillon sont séparées les unes des autres par des lignes sans aucun identifiant *TreeID*
4	Token	Segment de la transcription orthographique pris en 2 blancs ou un blanc et une signe de ponctuation
5	Word Span	
6	Word form	
7	Lemma	Les lemmes sont comme il est d'usage la forme pour les lexèmes invariables, la forme infinitive pour les verbes, le singulier pour les noms et le masculin singulier pour les adjectifs.
8	POS	Partie du discours - V pour les verbes - N pour les noms - Adj pour les adjectifs - Adv pour les adverbes - Pre pour les prépositions - CS pour les conjonctions de subordination - J pour les joncteurs : il s'agit des traditionnelles conjonctions de coordinations et d'autres éléments qui lient les couches d'un entassement, comme *c'est-à-dire* ou *y compris*. Les éléments clôtureurs d'entassement comme *et caetera* sont classés comme joncteurs également. - D pour les déterminants - I pour les interjections, y compris des marqueurs de discours comme *bon, ben, euh, hein* ... - Qu pour les mots qu- que sont les relatifs et les interrogatifs - Cl pour les clitiques, y compris les clitiques sujets (*je, tu, il, on, ce*) et l'adverbe de négation *ne*. - Pro pour les autres pronoms - X pour les éléments dont on ne peut déterminer la catégorie syntaxique : partie inaudible (XXX), certaines amorces (quand on ne peut pas deviner le lexème et sa partie du discours), ainsi que les positions non instanciées marquées par &.
9	Mood	
10	Tense	Seuls les V à l'indicatif varient en temps ; le trait *tense* possède 5 valeurs : *present, imperfect, future, conditional* et *perfect*
11	Person	Les V reçoivent aussi des traits d'accord : le trait *person* a trois valeurs *1, 2* et *3*
12	Number	le trait *number* a deux valeurs *sg* et *pl*
13	Gender	le trait *genre* a deux valeurs *fem* et *masc*
14	ID dep	
15	Type dep	
16	ID plain	
17	Type plain	
18	ID junc	
19	Type junc	
20	ID para	
21	Type para	
22	ID inherited	
23	Type inherited	
24	ID junc_inherited	
25	Type junc_inherited	
26	Layer	
27	UI	
28	Nucleus	
29	Prenucleus	
30	Gov nucleus	
31	Innucleus	
32	Gov Inncleus	
33	Postnucleus	
34	Gov postnucleus	

35	IU parenthesis	
36	IU graft	
37	IU embedded	
38	Associated nucleus	
39	Intro UI	
40	Prominence final	Proéminence de la syllable finale du mot courant
41	Prominence initial	Proéminence de la syllable initiale
42	Hesitation	Présence d'une hesitation
43	Pitch avg	Valeur moyenne de la hauteur
44	Pitch	Hauteur
45	Syllabe length	Longueur de la syllable du mot courant
46	Syllabe length avg	Longueur moyenne de la syllable du mot courant
47	Speaker	Locuteurs
48	Pause length	Longueur de la pause

(description à compléter cf *Rhapsodie*)

Les annotations 14 à 25 sont réutilisées 2 à 2 (*cf* jeu de couleur) pour construire respectivement une seule annotation (de type relation) qui est réécrite par exemple sous la forme :

`Type_dep(ID_dep)`

pour les lignes 14 et 15.

Les autres annotations sont réutilisées telles quelles par le processus de transcodage.

3. Intégration des annotations Rhapsodie dans une base textométrique

Les données issues de *Rhapsodie* transcodées dans un format compatible avec le *Trameur* respectent la structuration d'une base textométrique (*cf* documentation *Trameur*). Celle-ci est composée de 2 parties permettant :

1. La description d'une *Trame* textométrique : liste des items numérotés et annotés (ici chaque item est associé à 38 annotations)

2. La description du *Cadre* textométrique : liste des partitions définies sur la trame ; chacune porte un nom et est associée à une liste de parties définies chacune par son nom (le nom de l'échantillon de *Rhapsodie*), par sa position de début sur la *Trame* et sa position de fin

Le fichier issu du transcodage est au final une base textométrique qui se charge dans le *Trameur* par le module « *importation de base* ».

3.1 La Trame textométrique

La *Trame* textométrique définit la liste des items qui la composent :

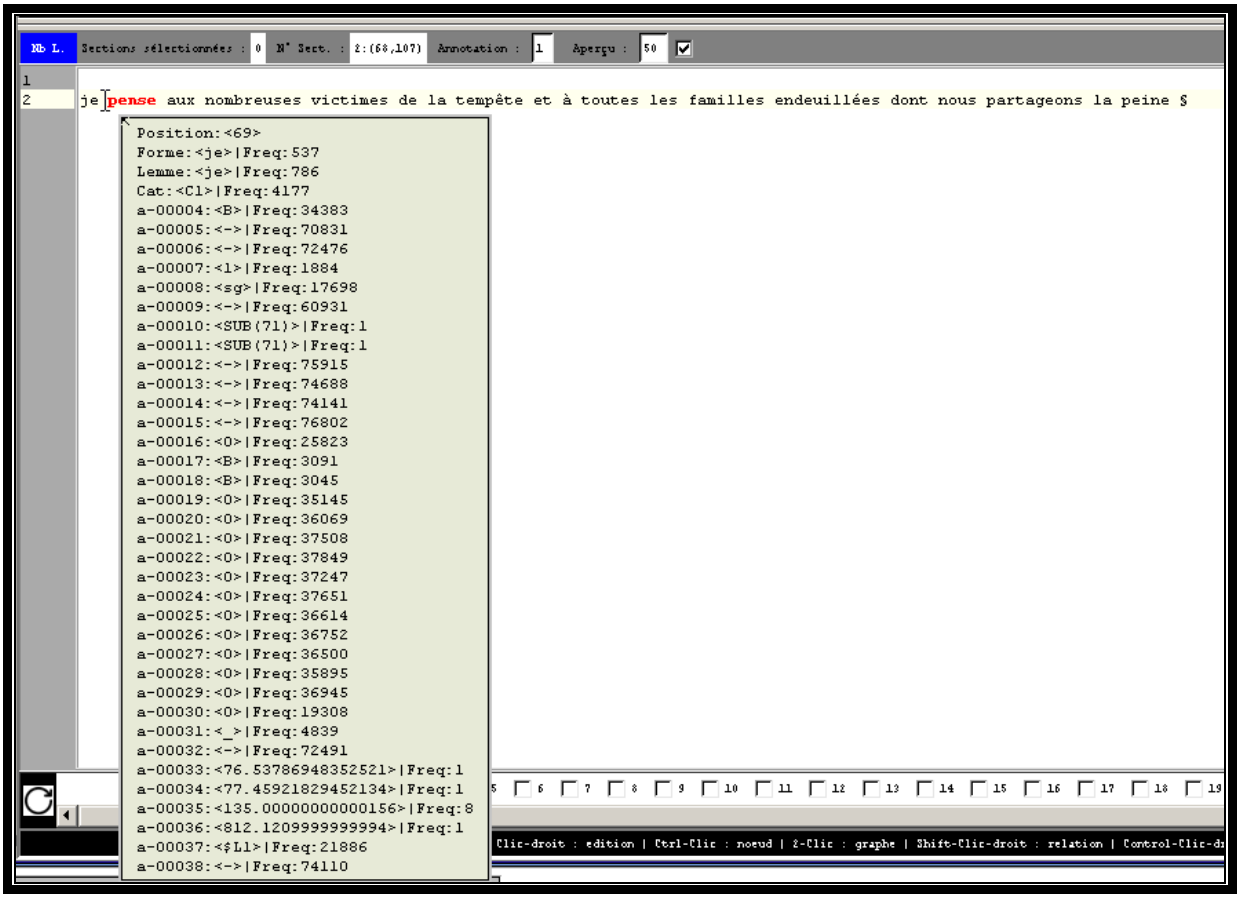
Chaque `token` de *Rhapsodie* devient un *item* de la *Trame*, il est identifié par sa position sur la *Trame* : l'identification des `tokens` définie dans les relations de dépendance fournies au départ dans *Rhapsodie*, est donc, après transcodage, la position de l'item dans la base.

Toutes les annotations initiales sont associées à chacun des items de la *Trame*.

Les annotations de dépendance sont fusionnées deux à deux sous la forme d'une relation marquée de la manière suivante : `RELATION(CIBLE)` (*cf* documentation *Trameur*)

- `RELATION` est une chaîne portant le nom de la relation visée
- `CIBLE` est une valeur numérique pointant vers une position de la *Trame*

Chaque item est associé *in fine* à 38 annotations, la figure suivante donne à voir ces annotations sur un item en contexte :



n°Annotation	Label (annotation trameur)	Contenu Rhapsodie
1	Forme	Token
2	Lemme	Lemma
3	Cat	POS
4	a-00004	Word_span
5	a-00005	Mood
6	a-00006	Tense
7	a-00007	Person
8	a-00008	Number
9	a-00009	Gender
10	a-00010	Type_dep(ID_dep)
11	a-00011	Type_plain(ID_plain)
12	a-00012	Type_junc(ID_junc)
13	a-00013	Type_para(ID_para)
14	a-00014	Type_inherited(ID_inherited)
15	a-00015	Type_junc_inherited(ID_junc_inherited)
16	a-00016	Layer
17	a-00017	IU
18	a-00018	Nucleus
19	a-00019	Prenucleus
20	a-00020	Gov_nucleus
21	a-00021	Innucleus
22	a-00022	Gov_innucleus
23	a-00023	Postnucleus
24	a-00024	Gov_postnucleus
25	a-00025	IU_parenthesis
26	a-00026	IU_graft
27	a-00027	IU_embedded
28	a-00028	Associated_nucleus
29	a-00029	Intro_IU
30	a-00030	Prominence_final
31	a-00031	Prominence_initial
32	a-00032	Hesitation
33	a-00033	Pitch_avg
34	a-00034	Pitch
35	a-00035	Syllabe_length
36	a-00036	Syllabe_length_avg
37	a-00037	Speaker
38	a-00038	Pause_length

3.2 Le Cadre textométrique

Les différents échantillons initiaux de *Rhapsodie* sont considérés comme autant de parties différentes : la base finale est donc une partition de textes (*Cadre*), chaque partie contient les zones textuelles associées à l'identifiant initial de l'échantillon.

(cf. **Annotations Rhapsodie pour le Trameur**)

3.3 Sections

Le processus de transcodage intègre aussi un marquage de sections : après chaque UI, un caractère délimiteur de section (§) est introduit pour permettre de construire dans le *Trameur* une représentation cartographique de la base sous la forme d'une *carte des sections*.

(cf. **Annotations Rhapsodie pour le Trameur**)

4. Explorer les relations de dépendance

Les différentes fonctionnalités disponibles dans le *Trameur* pour travailler avec les annotations de relations sont décrites dans la documentation du *Trameur*.

(cf. **Annotations Rhapsodie pour le Trameur**)