Projet FRFC " Périphérie gauche des unités de discours " - Protocole de codage syntaxique

Noalig Tanguy, Thomas van Damme, Liesbeth Degand, Anne-Catherine Simon

Tanguy Noalig, Van Damme Thomas, Degand Liesbeth & Simon Anne-Catherine

Université Catholique de Louvain – Institut Langage & Communication –
Centre VALIBEL – Discours & Variation
2012

Projet FRFC « Périphérie gauche des unités de discours »
Protocole de codage syntaxique

Le projet FRFC [convention FRFC 2.4524.11 (17468261)] portant sur la *Périphérie gauche des unités de discours : analyse syntaxique, prosodique et fonctionnelle et étude de l'impact sur la macro-organisation du discours* et mené au centre Valibel Discours et Variation de l'Université Catholique de Louvain étudient les différents enjeux suivants : (i) définition à la fois syntaxique et prosodique de la périphérie gauche / position initiale des unités discursives de base (BDU), (ii) établissement d'une typologie (paradigmes et fonctions) des éléments que l'on retrouve à l'initiale des BDU, (iii) analyse de la distribution des EPG selon les genres de discours et (iv) modélisation de la fonction de ces EPG au niveau de l'interprétation des textes.

Le premier objectif consiste en une catégorisation formelle et fonctionnelle des unités présentes en périphérie gauche de l'énoncé en vue de déterminer si celles-ci peuvent être considérées comme un paradigme discursif caractérisé par des traits syntaxiques, prosodiques et fonctionnels spécifiques. Pour cela, l'annotation syntaxique du corpus (données orales attestées d'une durée totale de 3h16) représente la première étape du travail.

Totalement indépendant du codage prosodique, le codage syntaxique a été réalisé manuellement sous Praat[1], en double annotation. Pour situer cette étape dans une optique quasi exclusivement syntaxique, nous avons écarté dans un premier temps tout recours à la prosodie et avons axé notre travail de segmentation à partir de la transcription orthographique du texte, en partant du principe que le repérage des unités syntaxiques d'un texte oral s'effectue par des critères syntaxiques et non à l'interface syntaxique / prosodie. Le recours au son a néanmoins été nécessaire pour définir le statut de certains éléments, notamment la portée gauche ou droite des séquences régies ou associées.

Les principes théoriques adoptés pour l'annotation syntaxique de ce projet font suite aux principes qui ont présidé pour le codage des productions orales du projet FRFC *Établissement d'une procédure de segmentation du discours oral en unités minimales sur base de l'interaction de critères syntaxiques et prosodiques*[2] (Degand et al. 2010 ; Degand & Simon 2005, 2009a & 2009b).

1. Procédure d'annotation syntaxique sous Praat

1.1. Méthode générale

Le travail de codage syntaxique réalisé sous Praat se présente sous six tires distinctes. Le fichier son du texte y est tout d'abord importé et apparaît dans la zone supérieure. Les six tires correspondent aux différentes étapes de l'annotation : (1) phonèmes, (2) syllabes, (3) mots, (4) rection, (5) séquences et (6) commentaires.

[1] Téléchargeable à l'adresse http://www.fon.hum.uva.nl/praat/.
[2] Projet financé par le FRS-FNRS de janvier 2007 à décembre 2008 et attribué à L. Degand et A-C Simon.

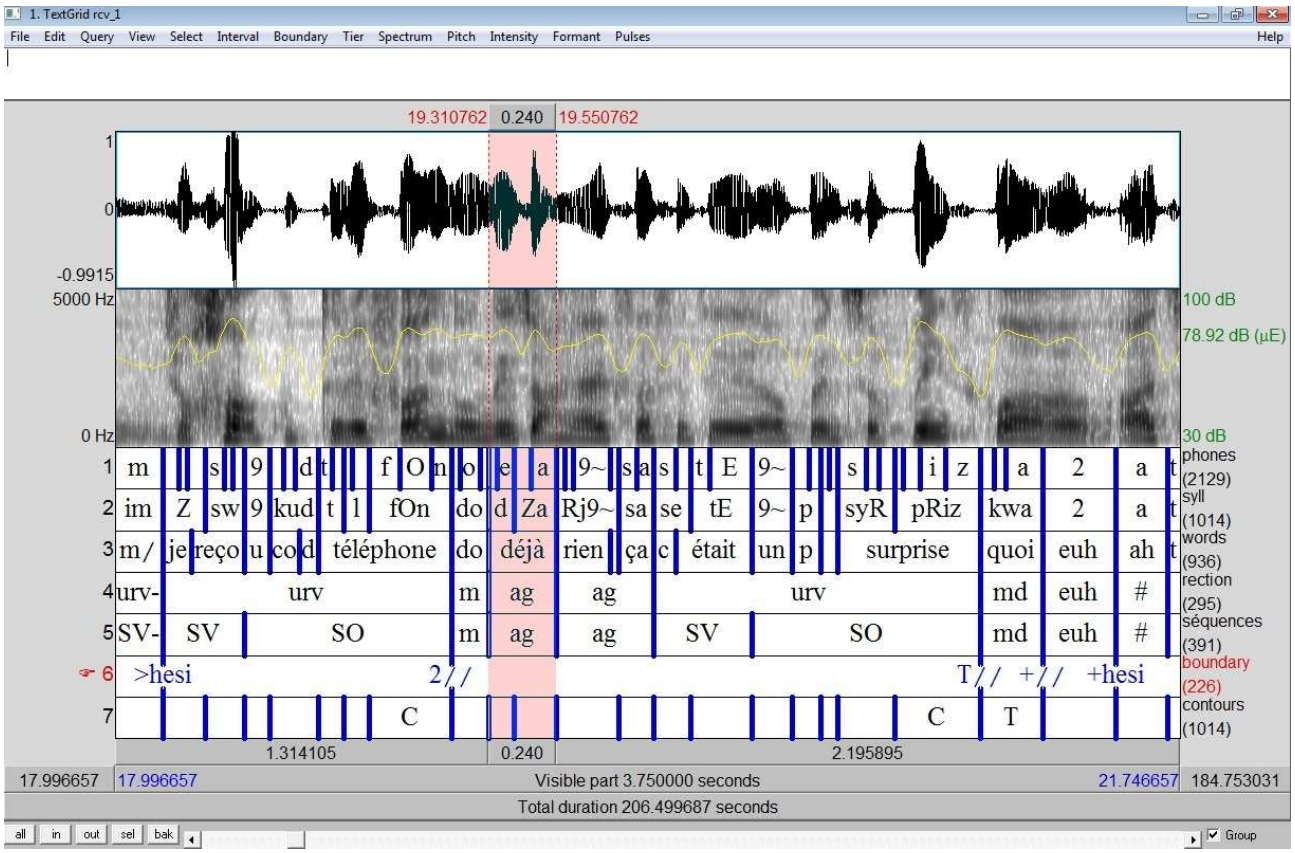

Figure 1 : Copie d'écran de la procédure d'annotation syntaxique sous Praat

1.2. Les tires d'annotation

1.2.1. Les tire « phonèmes », « syllabes » et « words »

Les extraits sonores sont tout d'abord alignés à l'aide du logiciel EasyAlign (Goldman 2010). Sur base d'une première transcription manuelle dans Praat en unités plus ou moins grandes (en moyenne des séquences de 2,5 secondes, selon la disposition des pauses silencieuses), le logiciel donne une première transcription phonétique de l'extrait, qui est corrigée ensuite manuellement (liaisons, enchaînements, prononciations « non prototypiques », etc.).

Dans un deuxième temps, le logiciel crée trois tires finales : (i) la première contient le texte segmenté en mots graphiques (« words »), (ii) la deuxième correspond à l'alignement en syllabes (« syll ») et (iii) la troisième un alignement en phonèmes (« phones »). Ce résultat est aussi par la suite validé manuellement.

Le codage syntaxique repose sur la tire « words ». Le codage prosodique, quant à lui, repose sur la tire « syll ».

1.2.2. La tire « rection »

La tire « rection » segmente le texte en unités de rection (ur). Toute unité rectionnelle est organisée autour d'un noyau, c'est-à-dire un élément recteur. Il s'agit généralement un verbe tensé mais le noyau de l'unité peut également être un constituant averbal : nom, pronom, adverbe ou groupe prépositionnel prédicatif (Lefeuvre 1999). Ainsi, nous distinguons trois types d'unités de rection : (i) les unités de rection verbales (urv), (ii) les unités de rection averbales (ura) et (iii) les unités de rection elliptiques (ure). Ces trois types peuvent être précisés des mentions « inachevée » (ur*x*-I) et « plus » (ur*x*+), ce que résume le tableau ci-dessous.

Unité de rection	Code	Unité inachevée	Unité « plus »
Unité de rection verbale	urv	urv-I	urv+
Unité de rection averbale	ura	ura-I	ura+
Unité de rection elliptique	ure	ure-I	ure+

Tableau 1 : Inventaire des types d'unités de rection – tire « rection »

Nous avons également annoté dans la tire « rection » les marqueurs de discours (md), les adjoints (a), certains silences (sil) et certains « euh » (euh). Tous les éléments codés dans cette tire y sont notés en minuscule : urv, ura, ure, md, ur*x*-I, sil, a, euh, etc. L'élément « -I » y apparaît néanmoins en majuscule.

1.2.3. La tire « séquence »

La tire « séquence » segmente les unités de rection en unités inférieures : les séquences. Ces éléments y font l'objet d'une catégorisation, fonctionnelle ou catégorielle selon le type d'unité de rection relevé.

Le découpage fonctionnel s'inspire de Bilger et Campione (2002) et concerne les unités rectionnelles, verbales et elliptiques. Ainsi, nous avons dégagé quatre types de séquences : (i) la séquence sujet (SS), (ii) la séquence verbale (SV), (iii) la séquence objet (SO) et (iv) la séquence régie (SR).

Les unités rectionnelles averbales (cf. infra 2.3.) sont précisées en termes de catégories et non de fonctions : SN, SPron, SAdj, etc.

Les unités rectionnelles elliptiques, selon les cas, sont précisées en séquences fonctionnelles ou catégorielles (cf. infra 2.4.).

Enfin, nous doublons dans cette tire les marqueurs de discours (md), les adjoints (a), certains silences (sil), certains « euh » (euh) (cf. supra figure 1) et les éléments paraverbaux (#).

Par ailleurs, les séquences régies (SR) et les adjoints (a), selon leur place par rapport à l'unité de rection ou le verbe auxquels ils se rapportent, sont précisés des mentions « g » et « d » en minuscules. Tous les autres éléments codés dans la tire d'annotation « séquences » y sont notés en majuscules (sauf l'insertion, les marqueurs de discours et les adjoints) : SS, SV, SO, SRg, SRd, ag, ad, insert, etc.

1.2.4. La tire « commentaires »

La tire « commentaires » permet aux différents codeurs de noter durant l'annotation leurs remarques, interrogations, hésitations, etc. Cette tire est supprimée une fois la segmentation validée.

2. Les unités de rection : vue d'ensemble

2.1. Tableau synthétique

Tire « rection »	**urv** *Unité de Rection Verbale*	**ura** *Unité de Rection Averbale*	**ure** *Unité de Rection Elliptique*
Tire « séquence »	SV (SS) (SO) (SRg ou SRd)	SN ou SPron ou SAdj ou SAdv ou SPrep ou SSub ou SInt ou SPart	(SS) et/ou (SO) et/ou (SRg ou SRd) ou SN, SPro, SAdj, S SAdv, SPrep, SSub, SInt, SPart
Remarques	Les parenthèses indiquent que les éléments ne sont pas obligatoires dans l'unité rectionnelle. Pour qu'un ensemble soit analysé comme une « urv », il faut au minimum une « SV ».		Dans une « ure », la « SV » est non exprimée. L'« ure » comporte alors au moins l'une des séquences : SS, SO ou SR.

Tableau 2 : Synthèse du codage des trois types d'unité de rection

Comme toute unité syntaxique, les unités de rection relevées sont de taille variable. Elles peuvent compter un seul mot comme plusieurs séquences très détaillées.

(1) [déchiffre]^{urv}

(2) [ouais]^{ura}

(3) [nous venons d'entendre les premières pages de votre nouveau livre L'Hymne]^{urv}

(4) [on dit que le jour où il apprit l'assassinat de Martin Luther-King il se trouvait dans un bar fréquenté par les blancs il garda un silence mortel lorsqu'une type gueula bon débarras que son visage resta de marbre lorsqu'un autre se mit à rugir c'est une bonne leçon pour les nègres qu'il versa très lentement le sucre dans son café lorsque le barman avec un affreuse expression de joie commenta bien fait le bamboula l'a bien cherché qu'il avala très lentement sa boisson malgré les bonds que faisait son coeur serré comme le poing jusqu'à sa bouche qu'il refoula au fond de lui une colère veille de plusieurs siècles une colère héritée d'un peuple qui avait appris pour sauver ses billes à ne pas parler inconsidérément et que le lendemain de ce drame le cinq avril mille-neuf-cent-soixante-huit à New-Ark sur la scène du Symphony Hall il rendit un hommage inoubliable à l'homme assassiné et fit jaillir en beauté sauvage la douleur concentrée immobile et muette qu'il avait la veille u prix d'un effort inhumain contenu]^{urv} [3]

Par ailleurs, nous nous sommes réduits à segmenter uniquement le premier niveau syntaxique sans préciser la composition de certaines séquences (comme par exemple les subordonnées en fonction d'objet) qui comportent tout de même une structure d'unité de rection en SS, SV, etc. Ainsi, pour (5) :

(5) [(je pense)$_{SV}$ (que votre éditeur il préfère vous avoir que pas vous avoir)$_{SO}$]^{urv}

L'unité rectionnelle « *votre éditeur il préfère vous avoir que pas vous avoir* » fait partie de l'unité « *je pense que votre éditeur il préfère vous avoir que pas vous avoir* » en tant que séquence objet

[3] Les exemples (1) à (4) ne sont pas ici précisés au niveau séquentiel.

(SO) (autrement dit, COD). Cela concerne tous les éléments constitutifs de toute proposition subordonnée qui ne seront pas précisés dans la tire séquences. Le découpage en séquences se fait uniquement par rapport au verbe recteur de premier niveau syntaxique. Ainsi, pour (6) :

(6) [(il avait l'impression)$_{SV}$ (que quelque chose fouettait l'air près de lui)$_{SO}$]^{urv}

Le groupe nominal « *l'air* » est régi par le verbe « *fouetter* ». Cependant, il ne fait pas partie du premier niveau syntaxique de l'unité rectionnelle mais relève d'un plan subordonné : il appartient à la subordonnée complétive comme SO du SV « *fouettait* ». Le syntagme « *l'air* » (tout comme le groupe prépositionnel « *près de lui* », etc.) ne reçoit donc pas de découpage dans la tire séquence.

2.2. Les unités de rection verbale (codées « urv »)

Une unité de rection verbale possède nécessairement un verbe, ou du moins sous-entendu (l'unité est dans ce cas dite « elliptique »). Elle comprend également, mais pas systématiquement, les différents éléments régis par le verbe : les séquences sujet (SS), objet (SO), et régies (SR) :

(7) [(un phonostyle)$_{SS}$ (va être interprété)$_{SV}$ <euh> (par un auditeur)$_{SO}$]^{urv}

(8) [(j'ai dit)$_{SV}$ (tout à l'heure)$_{SRd}$ (une profession ou des choses comme ça)$_{SO}$]^{urv}

Dans notre étude, les verbes recteurs ne reçoivent pas de traitement particulier. Ainsi pour (9), nous considérons que la complétive « *que c'était un mois et demi deux mois après* » est régie par le verbe « *croire* ».

(9) [(je crois)$_{SV}$ (que c'était un mois et demi deux mois après)$_{SO}$]^{urv}

De la même manière, pour (10), nous analysons l'ensemble en deux unités rectionnelles, reliées entre elles au niveau macro-syntaxique.

(10) [(c'etait)$_{SV}$ (vendredi)$_{SO}$]^{urv} [(je crois)$_{SV}$]^{urv}

2.3. Les unités de rection averbale (codées « ura »)

Les énoncés averbaux se répartissent en deux grands types d'emplois (Tanguy 2009) :

(i) Nous entendons par « unités de rection averbales » les structures prédicatives averbales complètes :

(11) [(oui)$_{SAdv}$]^{ura}

(12) [(le quartier Croydon)$_{SN}$ (sans doute le quartier le plus touché par les émeutes)$_{SN}$]^{ura}

(13) [(insupportable pour nos sociétés)$_{SAdj}$]^{ura}

(14) [(d' abord)$_{SRg}$ (l'augmentation des salaires et des pensions)$_{SN}$]^{ura}

Il existe en effet pour le français trois types de phrases averbales (Lefeuvre 1999 ; Tanguy 2009) : (a) les phrases à deux termes (sujet / prédicat) – exemple (12) –, (b) les phrases à un terme où seul le prédicat est formulé (Le sujet y est implicite. Son référent est présent dans le contexte situationnel ou linguistique) – exemples (11) et (13) –, et (c) les phrases existentielles ne comportant pas de sujet et marquant une existence – exemple (14).

(ii) Les constructions ne comportant pas cette structure mais étant tout de même interprétables comme des prédications (c'est-à-dire des unités de rection verbales) par recours au contexte sont analysées comme des unités de rection elliptiques (ure).

(15) L2 je sais pas pourquoi je te parle de ça d'ailleurs non mais voilà euh dans le sens être déjà grand
L1 [(**pour faire bon garçon**)$_{SPrep}$]^{ure}

(16) L1 : donc c'est l'histoire d'un personnage qui s'appelle Polza
L2 ouais [(**Polza Moncini**)$_{SN}$]^{ure}

Les prédications averbales, comportant un prédicat averbal et non verbal, ne peuvent être précisées en séquences de type fonctionnel : SS, SO, etc. Nous précisons donc uniquement la catégorie des séquences prédicat et/ ou sujet.

(17) [(hasard du calandrier)$_{SN}$]^{ura}

(18) [(comment)$_{SAdv}$ (ça)$_{SPron}$]^{ura}

(19) [(facile pour la mémoire)$_{SAdj}$]^{ura}

(20) [(peut- être)$_{SAdv}$ (que c' est par là qu' il faut que je regarde)$_{SSub}$]^{ura}

(21) [(chez larousse leurron)$_{SPrep}$]^{ura}

(22) [(aïe aïe aïe)$_{SInt}$]^{ura}

Nous regroupons en une seule unité de rection averbale une suite de plusieurs *oui / ouais / non* :

(23) [(oui oui)$_{SAdv}$]^{ura}

Cependant, l'extrait (24) :

(24) [(c'est ça)$_{SV}$]^{urv} [(oui)$_{SAdv}$]^{ura}

est analysé en deux unités de rection : une unité verbale, suivie d'une unité averbale.

2.4. Les unités de rection elliptique (notées « ure »)

Nous appelons unités de rection elliptique les unités incomplètes, mais interprétables comme des unités de rection verbales par recours au contexte. Cela concerne les réponses aux questions, les énoncés thématiques, les reprises, etc.

(25) L1 vous n'employez pas de dictionnaire de prononciation
L2 [(non)$_{SAdv}$]^{ura} [(**aucun dictionnaire de prononciation**)$_{SN}$]^{ure}

(26) L2 mais euh en tout cas euh il y a beaucoup de gens qui doivent venir te saluer à chaque fois
L1 <et>$_{md}$ [(**pas toi**)$_{SPron}$]^{ure}

Ces exemples sont à distinguer des unités comportant, au sein d'une même séquence, une coordination de constituants :

(27) [(il y a)$_{SV}$ (**les proustiens et les céliniens**)$_{SO}$]^{urv}

L'ensemble postverbal constitue une seule et même séquence objet (séquence d'une construction impersonnelle) et non deux 'compléments' coordonnés : impossibilité d'extraire ou de pronominaliser un seul des éléments.

2.5. Compléments des unités de rection

2.5.1. Tableau synthétique

Tire « rection »	**…-I** *unité inachevée*	**…+** *unité « plus »*
Tire « séquence »	/	/
Remarques	Lorsqu'une séquence est jugée inachevée, elle doit aussi être spécifiée dans la tire « séquence » : SV-I, SS-I, SO-I ou SR-I.	**= ur contenant un adjoint, un insert ou un marqueur de discours intégré** L'élément inséré n'est pas codé dans la tire « rection », mais uniquement spécifié dans la tire « séquence ».

Tableau 3 : Synthèse du codage des compléments d'unités de rection

2.5.2. Les unités de rection inachevées (précisées par « … -I »)

Une unité de rection est dite « inachevée », syntaxiquement et sémantiquement, lorsqu'au moins l'un de ses compléments obligatoires y est absent – exemple (28) – et/ou lorsqu'une séquence amorcée est incomplète – exemple (29).

(28) $[(\text{on m'a dit})_{SV}\ \text{<euh>}]^{urv\text{-}I}$

(29) $[(\text{je crois})_{SV}\ (\text{qu'il})_{SO\text{-}I}]^{urv\text{-}I}$

Pour signaler l'inachèvement, nous ajoutons le signe « -I » au code initial de l'unité de rection. Par exemple : « urv-I » pour une unité rectionnelle verbale inachevée, « SO-I » pour une séquence objet inachevée, etc. Les unités de rection sans cette mention sont par défaut des unités complètes.

2.5.3. Les unités de rections « plus » (précisées par « … + »)

Toute unité de rection, qu'elle soit verbale, averbale ou elliptique, peut insérer un adjoint, un insert ou un marqueur de discours. Ces éléments n'entrent pas dans la rection du verbe même s'ils sont insérés dans l'unité rectionnelle. Lorsqu'une ur contient au moins l'un de ces trois éléments, celle-ci est précisée par le signe « + ».

(30) $[(\text{j'en remercie})_{SV}\ \text{<}\mathbf{\text{d'ailleurs}}\text{>}_{md}\ \mathbf{\text{<cher Hervé>}}_{\mathbf{ad}}\ \mathbf{\text{<cher Alain>}}_{\mathbf{ad}}\ (\text{tout spécialement})_{SRd}\ (\text{les ministres de la défense qui sont parmi nous ce soir})_{SO}]^{urv+}$

(31) $[(\text{on peut dire})_{SV}\ \text{<}\mathbf{\text{je pense}}\text{>}_{insert}\ (\text{une star})_{SO}]^{\mathbf{urv+}}$

3. Les éléments non rectionnels

3.1. Tableau synthétique

Tire « rection »	**md** *marqueur de discours*	**a** (**ag** ou **ad**) *adjoint*	**euh**	**sil** *silence > 250 ms*	**#**	***non codé*** *insertion*	***non codé*** *marqueur d'interrogation*
Tire « séquence »	md	a (ag ou ad)	euh	sil	#	insert	mi
Remarques		L'adjoint est intégré ou périphérique à l'ur à laquelle il se rattache.		Si le silence est inférieur à 250 ms, il doit être rattaché à ce qui suit.		**= parenthèse**	**= « est-ce qu- »**

Tableau 4 : Synthèse du codage des éléments non rectionnels

Lorsque les divers éléments – marqueurs de discours, adjoints, *euh*, silences et insertions – apparaissent entre deux unités de rection, ils sont codés, dans la tire « rection », dans un intervalle séparé. L'intervalle est dupliqué dans la tire « séquence ». Ainsi, l'élément est annoté à deux reprises de manière identique :

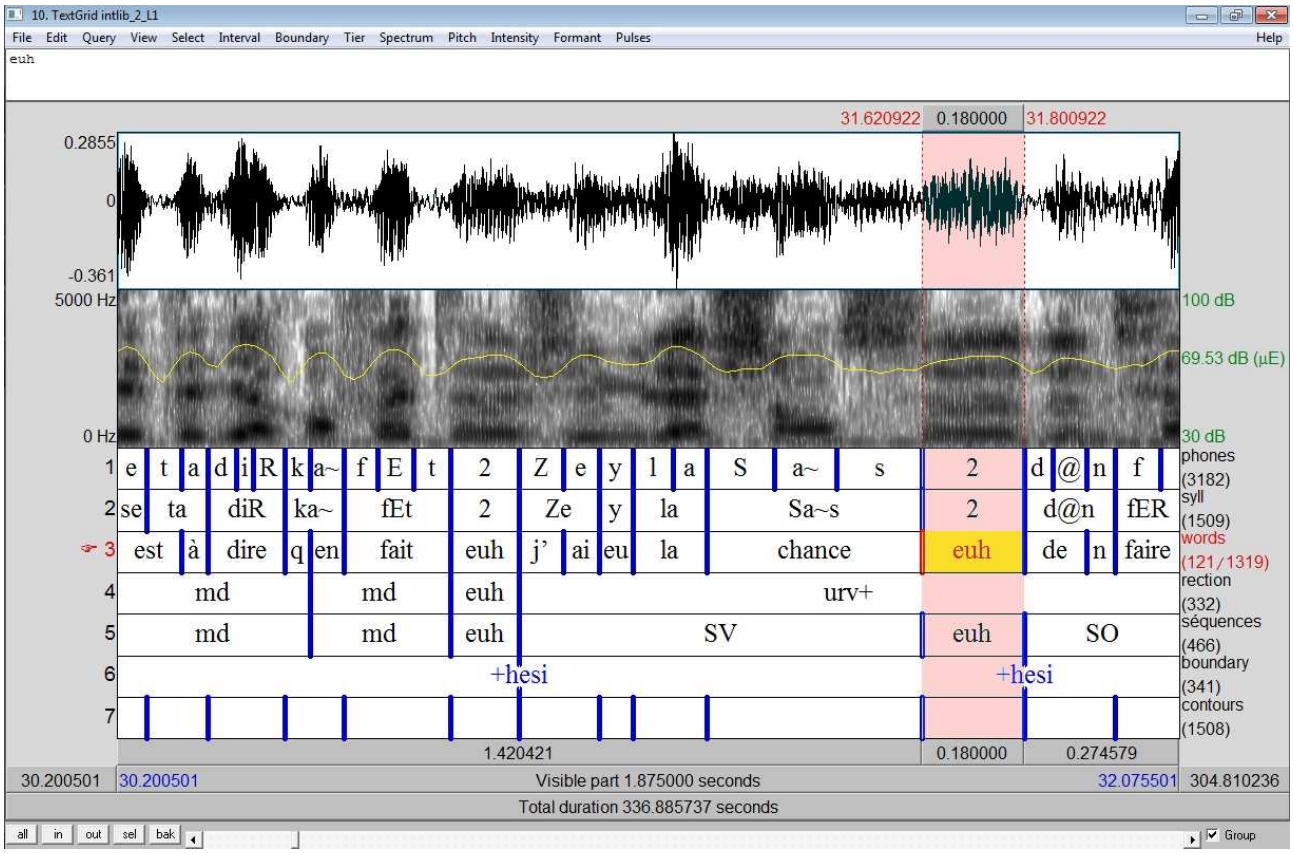

Figure 2 : Annotation des éléments non rectionnels

(32) [(je m'y attendais pas du tout)$_{SV}$]^{urv} **<euh>**$_{euh}$^{euh} (sil695ms) **<et>**$_{md}$^{md} [(c'est comme ça que)$_{SRg}$ (je suis devenue marraine)$_{SV}$]^{urv}

Cependant, si l'un de ces cinq éléments apparaît à l'intérieur d'une unité de rection, entre deux séquences, il est uniquement codé au niveau de la séquence.

(33) <mais>$_{md}^{md}$ <bon>$_{md}^{md}$ [(ça dure)$_{SV}$ **<euh>$_{euh}$** (sil 561ms) (dix secondes trente secondes une minute maximum)$_{SO}$]^{urv}

Les silences, les *euh* et les marqueurs du discours apparaissant à l'intérieur même d'une séquence ne sont pas codés.

(34) [(c'est)$_{SV}$ <euh>$_{euh}$ (un petit dessin avec **<euh>** avec un petit bébé)$_{SO}$]^{urv} <quoi>$_{md}$

3.2. Les marqueurs de discours (codés « md »)

Les marqueurs de discours (md) marquent une relation entre deux unités de rection. Nous regroupons sous l'appellation « marqueurs de discours » aussi bien les connecteurs (ou connecteurs textuels), les conjonctions (de coordination et de subordination) que les marqueurs de discours à proprement parler, c'est-à-dire les autres lexèmes à fonction structurante et/ou métadiscursive. Un marqueur de discours n'est jamais régi (ce qui permet de distinguer certains adverbes des marqueurs de discours).

Ainsi, nous avons catégorisé comme marqueurs de discours les différents éléments :

ainsi	cependant	également	néanmoins	pourtant
alors	d'ailleurs	en effet	notamment	puis / pis
après	de ce fait	en fait	or	quand même
au contraire	de même	en plus	ou	quoi
au fond	de plus	enfin	oui / ouais / non	sinon
aussi	déjà	ensuite	par ailleurs	tiens
ben / bon	donc	et	par conséquent	toutefois
car	du coup	mais	par contre	tu vois

Tableau 5 : Liste des principaux marqueurs de discours codés

Les marqueurs de discours sont codés de la même manière, que ce soit dans la tire « rection » ou dans la tire « séquence », par le signe « md ». Cependant, ils sont indiqués dans la tire « rection » uniquement lorsqu'ils apparaissent à la périphérie de l'unité de rection. Le signe « + » vient compléter l'ur lorsqu'un marqueur est inséré dans l'unité (et uniquement codé dans la tire « séquence »).

3.3. Les adjoints (codés « a »)

L'adjoint (ou « associé ») se présente comme un élément qui n'entre pas dans la rection du verbe mais qui lui est toutefois associé. Il est dit « périphérique » à l'unité de rection à laquelle il se rattache. Selon leur place par rapport à l'unité rectionnelle, les adjoints sont précisés des mentions « g » et « d », en minuscule.

(35) **<s'il y a une illustration de la béatitude des coeurs purs>$_{ag}$** [(c'est bien)$_{SV}$ (Bernadette)$_{SO}$]^{urv}

(36) <mais>$_{md}$ <vous savez>$_{md}$ [(quand on est écrivain et qu'on fait de la littérature)$_{SRg}$ (on est grossier)$_{SV}$]^{urv} **<contrairement à ce que souvent on pense>$_{ad}$**

Cependant, lorsque l'élément associé apparaît à l'intérieur d'une séquence verbale, sa position est précisée par rapport au verbe ou l'auxiliaire verbal :

(37) [(j'en remercie)$_{SV}$ <d'ailleurs>$_{md}$ **<cher Hervé>$_{ad}$** **<cher Alain>$_{ad}$** (tout spécialement)$_{SRd}$ (les ministres de la défense qui sont parmi nous ce soir)$_{SO}$]$^{urv+}$

Le signe « + » vient alors compléter l'ur.

Hors de la rection verbale, les éléments associés ne peuvent être ni pronominalisés, ni entrer dans un procédé d'extraction :

(38) a. <donc>$_{md}$ **<évidemment>$_{ag}$** [(un phonostyle)$_{SS}$ (va être interprété)$_{SV}$ <euh> (par un auditeur)$_{SO}$]^{urv}

b. *donc c'est évidemment qu'un phonostyle va être interprété euh par un auditeur

La distinction adjoint (associé) / élément régi est cependant parfois difficile à opérer pour certains constituants.

Nous avons relevé comme éléments associés (adjoints) : les dislocations – exemples (39) et (40), les subordonnées *en puisque* – exemple (41), les marqueurs de point de vue – exemple (42), etc.

(39) **<la mobilité>$_{ag}$** [(nous connaissons)$_{SV}$]^{urv} <donc>$_{md}$

(40) <et>$_{md}$ <encore>$_{md}$ [(ça baisse)$_{SV}$ (un peu)$_{SRd}$]^{urv} **<la crédibilité>$_{ad}$**

(41) [(non)$_{SAdv}$]^{ura} **<puisque ça m'empêche pas de de tomber <euh> parfois dans des dépressions extrêmes>$_{ad}$**

(42) [(pourquoi)$_{SRg}$ (on vit)$_{SV}$ (une sale époque)$_{SO}$]^{urv} **<selon vous>$_{ad}$**

3.4. Les morphèmes d'hésitation « euh » (codés « euh »)

Les morphèmes d'hésitation « euh » sont indiqués dans la tire « rection » et/ou dans la tire « séquence » uniquement s'ils apparaissent entre deux unités de rection ou entre deux séquences ou autre élément. Par conséquent, nous ne relevons pas les « euh » à l'intérieur des séquences (cf. supra 3.1.).

3.5. Les silences (codés « sil »)

Nous avons dissocié trois codages de silences.

	silence < 250 ms	**silence > 250 ms entre deux ur**	**silence > 250 ms dans une ur**
Tire « rection »	*non codé*	sil	*non codé*
Tire « séquence »	*non codé*	sil	sil

Tableau 6 : Synthèse du codage des silences

Les silences supérieurs à 250 ms sont indiqués dans la tire « rection » et/ou dans la tire « séquence » s'ils apparaissent entre deux unités de rection ou entre deux séquences ou autre élément.

(43) [(oui)$_{SAdv}$)]^{ura} **(sil 550ms)** [(je pense)$_{SV}$ (que l'hymne est une moment infiniment plus grand qu'un moment muscial)$_{SO}$]^{urv}

(44) [(vous préconisez)$_{SV}$ <effectivement>$_{md}$ **(sil 313ms)** (l'impôt sur le revenu à la source)$_{SO}$]$^{urv+}$

Les silences inférieurs à 250 ms n'apparaissent pas dans un intervalle séparé. Ils sont rattachés à l'intervalle qui suit : unité de rection ou séquence.

(45) [(la diversité des cultures de notre pays)$_{SS}$ (sil 1500ms) (si elle est (sil 527ms) bien vécue)$_{SRg}$ (sil 1360ms) (constitue)$_{SV}$ (**(sil 237ms)** un formidable atout)$_{SO}$]^{urv}

3.6. Les éléments paraverbaux (codés « # »)

Nous avons rassemblé sous l'étiquette « paraverbale » les différents éléments : *hé, mh, hm, mm* – exemple (46), les sons non identifiés, ainsi que les rires – exemple (47), ricanement, toux, soupirs, etc.

(46) [(c'est très autobiographique)$_{SV}$]^{urv} **<mm>**$_{\#}$ <et>$_{md}$ [(c'est)$_{SV}$ <et>$_{md}$ (c'est assez passionnant)$_{SV\text{-}S}$ (parce que ça se lit)$_{SRd}$ <effectivement>$_{md}$ <quelqu'un vous l'a dit d'ailleurs>$_{insert}$ (comme un roman)$_{SRd\text{-}S}$]$^{urv+}$

(47) [(je sais très bien)$_{SV}$ (que ça n'existe pas)$_{SO}$]^{urv} **<rire>**$_{\#}$ [(comment)$_{SRg}$ (je peux vous expliquer)$_{SV}$ (ça)$_{SO}$]^{urv}

Le signe « # » apparaît à la fois dans la tire « rection » et dans la tire « séquence ».

3.7. Les insertions (codées « insert »)

Nous appelons « insertion » toute unité de rection insérée dans une autre unité de rection. Il s'agit très souvent de parenthèses – exemple (48), d'incises – exemples (49) et (50), d'unité à recteur faible – exemple (50) :

(48) [(nous voulions) <**et c' est la raison pour laquelle du choix de la date de ce colloque**>$_{insert}$ (profiter)$_{SV\text{-}S}$ (de la présence de Deirdre Wilson qui a été honorée comme docteur honoris causa hier)SO]$^{urv+}$

(49) [(l' identité)$_{SS}$ (est)$_{SV}$ <**on l' a rappelé**>$_{insert}$ (un effet structurel un rapport)$_{SO}$]$^{urv+}$

(50) [(il a)$_{SV}$ <quand même>$_{md}$ (accroché)$_{SV\text{-}S}$ (sa souffrance)$_{SO}$ <**disait Jean-Daniel>**$_{insert}$ (à une caste à un truc à une race)$_{SO}$]$^{urv+}$

(51) [(on peut dire)$_{SV}$ <**je pense**>$_{insert}$ (une star)$_{SO}$]$^{urv+}$

Au niveau du codage, l'insertion est uniquement codée dans la tire « séquence » par le code « insert ». Par conséquent, l'unité n'est pas décrite en SS, SV SO, etc. Le signe « + » vient compléter l'ur. Par ailleurs, si l'insertion scinde une séquence en deux intervalles, la seconde comprend le signe « -S » qui indique que la séquence correspond à la suite d'une première séquence déjà énoncée – exemple (48).

3.8. Le marqueur d'interrogation « est-ce qu- » (codé « mi »)

Le morphème spécifique (particule interrogative) *est-ce qu-* reçoit un codage spécifique comme marqueur d'interrogation « mi » dans la tire « séquence » uniquement :

(52) [<**est-ce que**>$_{mi}$ (vous croyez)$_{SV}$ (que jusqu'à présent notre conversation est suffisamment élevée intelligent inté/ intéressante)$_{SO}$]^{urv}

(53) <donc>$_{md}$ [(qu')$_{SS}$ <**est-ce qui**>$_{mi}$ (est plus facile qu'avant)$_{SV}$]^{urv}

4. Les séquences

La tire séquence segmente les unités rectionnelles en séquences de type fonctionnel ou catégoriel, selon le type d'unité rectionnelle (verbale, averbale ou elliptique).

4.1. Les séquences de type fonctionnel

Le découpage fonctionnel s'inspire des travaux de Bilger et Campione (2002). Il concerne les unités rectionnelles verbales et elliptiques.

Nous distinguons quatre types de séquences : (i) la séquence sujet (SS), (ii) la séquence verbale (SV), (iii) la séquence objet (SO) et la séquence régie (SR). Il s'agit ici de séquences maximales « qui représentent les constituants fonctionnels rencontrés dans les textes sans entrer dans le détail de leur composition » (Bilger et Campione 2002 : 118).

4.1.1. La séquence sujet (codée « SS »)

Un sujet est catégorisé comme une séquence sujet lorsqu'il apparaît sous une forme lexicale :

(54) [(monsieur Barnier)$_{SS}$ (nous annonce)$_{SV}$ (une commission d'enquête carrément sur le prix de la salade)$_{SO}$]^{urv}

Par conséquent, les sujets clitiques sont rattachés au verbe et font partie de la séquence verbale :

(55) <en tout cas>$_{md}$ [(j'ai peur)$_{SV}$ (du malheur)$_{SO}$]^{urv}

4.1.2. La séquence objet (codée « SO »)

Tout comme le sujet, tout objet est catégorisé comme une séquence à part entière lorsqu'il est réalisé par une forme lexicale pleine.

(56) [(mes élèves)$_{SS}$ (disent)$_{SV}$ (« sauf »)$_{SO}$]^{urv}

(57) <mais>$_{md}$ [(je le crois)$_{SV}$]^{urv}

L'étiquette « séquence objet » inclut aussi bien les compléments directs – exemple (58), les compléments indirects – exemple (59), dont les locatifs – exemple (60), les compléments d'agent – exemple (61), les attributs du sujet nominaux déterminés (SN définis et indéfinis) – exemples (62) et (63) et les séquences de constructions impersonnelles – exemple (64).

(58) [(si nous avons la responsabilité du pays)$_{SRg}$ (nous donnerons)$_{SV}$ (**des papiers**)$_{SO}$ (à tous ceux qui n'en ont pas)$_{SO}$]^{urv}

(59) [(cette demande que notre recteur adressera dans quelques instants à Madame la Ministre)$_{SS}$ (pourrait certes permettre de rendre)$_{SV}$ (un peu de cohérence)$_{SO}$ (**à un système de financement actuellement absurde**)$_{SO}$]^{urv}

(60) [(on est)$_{SV}$ (**dans un salon de coiffure**)$_{SO}$]^{urv}

(61) <et>$_{md}$ [(au niveau phonétique)$_{SRg}$ (ces accents primaires)$_{SS}$ (sont sou/ peuvent être réalisés)$_{SV}$ (**par des mouvements mélodiques qui peuvent être ou montants descendants**)$_{SO}$]^{urv} <puisque c'est les continuatifs mineurs de Delattre>$_{ad}$

(62) [<est-ce que>$_{mi}$ (le cinéma)$_{SS}$ (à sa naissance)$_{SRg}$ (était)$_{SV}$ (**un art magnifique qui a été euh ensuite euh détruit par le commerce**)$_{SO}$]^{urv}

(63) [(c'est)$_{SV}$ (**le Robert**)$_{SO}$]^{urv}

(64) <donc>$_{md}$ [(à partir de ces représentations métriques)$_{SRg}$ (on peut il peut y avoir)$_{SV}$ (**des accents initiaux**)$_{SO}$]^{urv}

4.1.3. La séquence verbe (codée « SV »)

La séquence verbale se compose du verbe et de ses différents actants clitiques et/ou pronominalisés – exemple (65). Elle peut également comporter les marques verbales de négation et de restriction – exemple (66), les modaux – exemple (67), les infinitifs sélectionnés par le verbe – exemple (68), certaines formes liées au verbe, figées ou non – exemples (69) et (70), les attributs adjectivaux – exemple (71) et les attributs nominaux non déterminés – exemple (72).

(65) <mais>$_{md}$ [(**je le crois**)$_{SV}$]^{urv}

(66) [(l'un)$_{SS}$ (**n'empêche pas**)$_{SV}$ (l'autre)$_{SO}$]^{urv}

(67) <donc>$_{md}$ [(le gouvernement)$_{SS}$ (**peut agir**)$_{SV}$]^{urv}

(68) [(**on peut ajouter**)$_{SV}$ (une détection automatique des syllabes proéminentes)$_{SO}$]^{urv}

(69) [(**il avait l'impression**)$_{SV}$ (que quelque chose fouettait l'air près de lui)$_{SO}$]^{urv}

(70) [(pourquoi)$_{SRg}$ <est-ce que>$_{mi}$ (la France)$_{SS}$ (**ferait une exception**)$_{SV}$]^{urv}

(71) [(**je suis célibataire**)$_{SV}$]^{urv}

(72) <et>$_{md}$ [(c'est comme ça que)$_{SRg}$ (**je suis devenue marraine**)$_{SV}$]^{urv}

4.1.4. La séquence régie (codée « SR »)

Une séquence régie relève de la rection verbale mais n'est pas un élément appelé syntaxiquement par le verbe (appartenant à la valence verbale). Elle se distingue d'une séquence associée en répondant favorablement aux tests de pronominalisation et d'extraction (Blanche-Benveniste et al. 1990).

(73) a. [(dans le voyage au bout de la nuit)$_{SRg}$ (il commence)$_{SV}$ (une ligne)$_{SO}$ (en disant)$_{SRd}$]^{urv}

b. c'est dans le voyage au bout de la nuit qu'il commence une ligne en disant

c. il y commence une ligne en disant

Tout comme les séquences associées, selon leur place (par rapport à l'unité rectionnelle ou le verbe), les séquences régies sont précisées dans la tire « séquence » des mentions « g » et « d », en minuscules.

(74) <finalement>$_{ag}$ [(**quand j'ai eu enfin déchiffré tout le truc**)$_{SRg}$ (je lui ai dit)$_{SV}$]^{urv}

(75) <mais>$_{md}$ [<est-ce que>$_{mi}$ (vous lisez)$_{SV}$ (**aussi mal que vous écrivez**)$_{SRd}$]^{urv} <Monsieur Dantzig>$_{ad}$

(76) [(**d'abord**)$_{SRg}$ (les salaires)$_{SN}$]^{ura}

De la même manière que la distinction associé / élément régi est parfois délicate, la séparation entre séquence régie et séquence objet à valeur locative n'est pas toujours évidente.

Enfin, lorsque plusieurs séquences régies se suivent, le codage les dissocie :

(77) [(cinq Palestiniens dont au moins trois membres de la branche armée du mouvement islamiste Hamas)$_{SS}$ (ont été tués)$_{SV}$ (**ce matin**)$_{SRd}$ (**dans des raids aériens israéliens et des échanges de tirs avec des soldats dans la bande de Gaza**)$_{SRd}$]^{urv}

4.2. Les séquences de type catégoriel

Les séquences de type catégoriel précisent les unités rectionnelles averbales et certaines unités rectionnelles elliptiques si celles-ci ne peuvent être « fonctionnalisées ».

4.2.1. Les séquences nominales (codées « SN »)

(78) [(plus six euros)$_{SN}$ (sur les retraites)$_{SRd}$]^{ura}

4.2.2. Les séquences adjectivales (codées « SAdj »)

(79) [(pas évident)$_{SAdj}$ (dans notre sociéte une sociéte où on se demande mais Dieu est-ce qu'Il existe)$_{SRd}$]^{ura}

(80) [(obèse)$_{SAdj}$]^{ura} [(ouais)$_{SAdv}$]^{ura}

4.2.3. Les séquences pronominales (codées « SPron »)

(81) <en tout cas>$_{md}$ [(moi)$_{SPron}$ (parce que j'avais choisi de vivre dans une famille d'accueil)$_{SRd}$]^{ure}

(82) [(comment)$_{SAdv}$ (ça)$_{SPron}$]^{ura} [(ça n'a)$_{SV}$ (aucune importance)$_{SO}$]^{urv}

4.2.4. Les séquences adverbiales (codées « SAdv »)

(83) [(exactement)$_{SAdv}$]^{ura} [(tout à fait)$_{SAdv}$]^{ura}

(84) [(oui)$_{SAdv}$]^{ura} [(allez-y)$_{SV}$]^{urv} [(non {non, ø})$_{SAdv}$]^{ura} [(allez-y)$_{SV}$]^{urv}

4.2.5. Les séquences prépositionnelles (codées « SPrep »)

(85) [(pour devenir quelqu'un)$_{Sprep}$]^{ure} [(pour avoir un papier)$_{Sprep}$]^{ure}

4.2.6. Les séquences infinitives (codées « SInf»)

(86) [(à suivre)$_{SInf}$]ure [(absolument)$_{SAdv}$]^{ura}

4.2.7. Les séquences subordonnées (codées « SSub »)

(87) [(bien sûr)$_{SAdj}$ (que nous ne serions pas socialistes si nous n'avions pas de l'optimisme de l'espérance)$_{Ssub}$]^{ura}

4.2.8. Les séquences interjectives (codées « SInt »)

(88) [(il klaxonne)$_{SV}$]^{urv} [(tut)$_{SInt}$]^{ura}

4.2.9. Les séquences participiales (codées « SPart »)

(89) [(photo de Chavez)$_{SN}$]^{ura} [(trapu)$_{SAdj}$]^{ura} [(les épaules ramassées)$_{SPart}$]^{ura}

4.3. Les séquences discontinues

Une séquence peut être interrompue par une autre séquence (insert, SR, associé, etc.) avant de se poursuivre. La suite de la séquence interrompue est codée par « -S » : SO-S, SV-S, SR-S, etc.

(90) <et>$_{md}$ [(c'est)$_{SV}$ (**un poème**)$_{SO}$ <au fond>$_{md}$ (**qui se lit comme une histoire qui se lit**)$_{SO-S}$ <je dirais>$_{insert}$ (**facilement euh qui mêle le réalisme le lyrisme**)$_{SO-S}$]$^{urv+}$

5. Export du codage syntaxique

Les unités de rection sont encadrées des crochets droits […] et précisées (urv, ura ou ure) à droite en exposant. Les séquences sont notées entre parenthèses (…) précisées à droite en indice.

(91) [(je m'y arrête)$_{SV}$ (un instant)$_{SRd}$]^{urv}

Les marqueurs de discours, adjoints (associés), inserts et marqueurs d'interrogation sont retranscrits entre crochets angulaires <…> précisés à droite en indice.

(92) [(on peut dire)$_{SV}$ <je pense>$_{insert}$ (une star)$_{SO}$]$^{urv+}$

Les hésitations sont également spécifiées entre crochets angulaires <…> : <euh>. Les éléments paraverbaux sont notés entre accolades et accompagnés du signe # en indice : {mm}$_{\#}$ / {rire}$_{\#}$. Les silences, ensuite intégrés au codage prosodique dans l'annotation des frontières, n'apparaissent plus dans le texte final.

6. Traitement des cas particuliers

6.1. Les reprises et les répétitions

Nous n'avons pas réservé de traitement particulier, dans notre codage, aux répétitions et aux cas de reprise lexicale ou de commutation dans une même position syntaxique :

(93) <et>$_{md}$ [(**je je** descends)$_{SV}$]^{urv}

(94) <maintenant>$_{md}$ [(**la deuxième rai/ le deuxième constat dont j'aimerais partir et qui n'est pas ici spécifique à une discipline mais qui est euh tout à fait général**)$_{SS}$ (concerne)$_{SV}$ <en fait>$_{md}$ (le type euh de formation ou les types de connaissance qu'on peut supposer devoir être celle(s) euh d'enseignants)$_{SO}$]^{urv}

6.2. Les séquences introduites par « c'est-à-dire qu- »

Les unités introduites par le tour *c'est-à-dire qu-* sont codées comme des unités de rection verbales et la forme figée est perçue comme un marqueur de discours.

(95) <mais>$_{md}$ <**c'est-à-dire qu'**>$_{md}$ <en fait>$_{md}$ <euh> [(j'ai eu la chance)$_{SV}$ <euh> (de ne faire que ça toute ma vie de pas avoir à chercher des boulots alimentaires)$_{SO}$]^{urv} <parce que nos parents qui étaient des des bons parents à mon frère à moi nous avaient payé un petit appartement en banlieue parisienne>$_{ad}$

6.3. Les séquences introduites par « ce qu- »

Les unités introduites par la forme *ce qu-* sont également codées comme des unités de rection verbales. Le sujet pronominal en *ce qu-* appartient alors à la séquence verbale.

(96) [(nous connaissons)$_{SV}$ (depuis longtemps)$_{SRd}$ (en Suisse)$_{SRd}$ (la diversité et la multiplicité des origines et des cultures)$_{SO}$]^{urv} [(**ce qui** ne veut pas dire)$_{SV}$ (que nous devrions abandonner nos particularités)$_{SO}$]^{urv} <au contraire>$_{md}$

6.4. Discours rapporté : discours direct

Traditionnellement, tout discours direct occupe la fonction de complément direct d'un verbe introducteur de discours (*dire, déclarer, demander,* etc.). Nous avons néanmoins préféré extraire les discours directs et les coder comme des unités de rection à part entière.

(97) [(Martine)$_{SS}$ (dit)$_{SV}$]^{urv} [(l'avenir)$_{SS}$ (aime)$_{SV}$ (la France)$_{SO}$]^{urv}

L'unité contenant le verbe introducteur n'est cependant pas codée comme inachevée.

6.5. La construction clivée, dispositif d'extraction

Le dispositif d'extraction tel que la forme clivée en *c'est ... qu-* (ou la variante en *il y a ... qu-*) reçoit dans notre codage un traitement particulier et n'est pas analysé comme une suite de SV-SO (Scappini 2006). Le constituant focalisé, encadré des éléments *c'est* et *qu-*, est codé selon la fonction qu'il occuperait dans un dispositif lié (forme canonique) :

(98) <car>$_{md}$ [(c'est Lui qui)$_{SS}$ (nous aime le premier)$_{SV}$]^{urv}

(99) <et>$_{md}$ [(c'est un des aspects qu')$_{SO}$ (on discutera)$_{SV}$ (dans la conclusion)$_{SRd}$]^{urv}

(100) <donc>$_{md}$ [(c'est pour ça que)$_{SRg}$ (je dis)$_{SV}$ (qu'il faut vraiment attendre l'ultime fin d'un écrivain pour le juger complètement)$_{SO}$]^{urv}

(101) <et>$_{md}$ <en fait>$_{md}$ [(on)$_{SV-I}$)]$^{urv-I}$ [(en plein milieu de la nuit)$_{SRg}$ (il y a eu un un A/ un Anglais qui)$_{SS}$ (a fumé)$_{SV}$ (dans le couloir)$_{SRd}$]^{urv}

6.6. Le dispositif pseudo-clivé en « ce qu- … c'est … »

Le dispositif pseudo-clivé, même s'il présente des similitudes avec le phénomène de la dislocation, fait également l'objet d'un traitement particulier (Blanche-Benveniste 2002 ; et al. 1984). La partie disloquée en *ce qu-* contient le verbe recteur, tandis que la seconde, introduite par la forme *c'est* comprend les différents éléments régis par le verbe SS ou SO. Notre codage aboutit aux découpages suivants :

(102) <mais>$_{md}$ [(ce qui était réconfortant d'un autre côté)$_{SV}$ (c'est qu'il y avait plein de gens dans cette euh dans dans cette maison)$_{SS}$]^{urv}

(103) <mais>$_{md}$ [(ce que je crois)$_{SV}$ (c'est qu'il faut d'entrée préparer un collectif budgétaire)$_{SO}$]^{urv}

Références bibliographiques

Bilger M. et Campione E. (2002). Propositions pour un étiquetage en "séquences fonctionnelles". *Recherches sur le français parlé* 17. pp. 117-136.

Blanche-Benveniste C. (2002). Phrase et construction verbale. *Verbum* XXIV, n°. 1: 23-36.

Blanche-Benveniste C., Deulofeu J., Stéfanini J. & Van den Eynde K. (1984). *Pronom et Syntaxe. L'approche pronominale et son application au français*. Paris : Selaf-Aelia.

Blanche-Benveniste C., Bilger M., Rouget C. & Van den Eynde K. (1990). *Le français parlé : études grammaticales.* Paris : Éditions du CNRS.

Degand L., Dister A. & Simon A. C. (2010). *Guide de codage pour le projet MDU. Partie syntaxique : découpage en unités de rection et en séquences fonctionnelles*. Rapport technique Centre de Recherche Valibel – Discours et Variation, 20 p.

Degand L. & Simon A.C. (2009a). *Mapping prosody and syntax as a strategic choice*. In: Wichmann A., Barth-Weingarten D. & Dehé N. (eds). *Where Prosody Meets Pragmatics*. Bangalore: Emerald. [Studies in Pragmatics, Volume 8], 79-105.

Degand L. & Simon A.C. (2009b). On identifying basic discourse units in speech: theoretical and empirical issues. *Discours* 4 (online-journal). [available online at URL:

http://discours.revues.org/index54.html]

Degand L. & Simon A.C. (2005). *Minimal Discourse Units: Can we define them, and why should we?* In: Aurnague M., Bras M., Le Draoulec A., & Vieu L. (eds). *Proceedings of SEM-05. Connectors, discourse framing and discourse structure: from corpus-based and experimental analyses to discourse theories,* Biarritz, 14-15 November 2005, 65-74. [available online http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=bras&subURL=sem05/proceedings-final/06-Degand-Simon.pdf]

Goldman J-P. (2010). *EasyAlign: a friendly automatic phonetic alignment tool under Praat*. University of Geneva. http://latlcui.unige.ch/phonetique/easyalign/easyalign_unpublished.pdf.

Lefeuvre F. (1999). *La phrase averbale en français*. Paris : L'Harmattan.

Scappini S-A. (2006). *Étude du dispositif d'extraction en « c'est…qu ». Différenciation entre une relative en « c'est…qu » et une proposition clivée*. Thèse de doctorat, Aix-en-Provence : Université de Provence.

Tanguy N. (2009). *Les segments averbaux, unités syntaxiques de l'oral*. Thèse de doctorat, Université Sorbonne Nouvelle Paris 3.