André Valli - | José Deulofeu ORFEO/TALEP

Guide de segmentation pour ORFEO

I. Principe de segmentation :

Les corpus oraux d'ORFEO se présentent de la manière suivante. La chaîne sonore est transcrite en orthographe standard, sans ponctuation ni annotation prosodique. Les éventuelles marques d'interrogation ou d'exclamation présentes dans les versions originales ainsi que les pauses ont été balisées en commentaires.

Notre objectif était de segmenter les transcriptions afin de permettre une analyse syntaxique en vue d'un traitement informatique. Le principe de cette segmentation est syntaxique : il s'agit de marquer le domaine à l'intérieur duquel s'établissent des relations de dépendance par rapport à un gouverneur. La segmentation adoptée permet de marquer les limites droite et gauche de l'énoncé.

Cette segmentation a été appliquée à la sous partie du corpus désignée par « corpus Gold », soit environ 250000 mots. La segmentation automatique sur le reste du corpus a été faite à partir d'un apprentissage sur la partie segmentée à la main dont les principes sont précisés dans ce guide.

La plupart des travaux de traitement automatique des langues considèrent la phrase comme une unité naturelle: mais la pertinence linguistique de cette notion fait pourtant pour le moins l'objet d'un débat (voir par exemple : Berrendonner, 2002 ; Blanche-Benveniste, 2002 ; Kleiber, 2003). On note d'ailleurs que même à l'écrit, phrases graphiques et unités linguistiques sont parfois non concordantes. La phrase est au mieux une approximation graphique, qui résulte d'un compromis entre structure syntaxique, intonation et mise en page. A l'oral, les majuscules et les points n'existent pas, et la notion de phrase y est encore moins opératoire.

Nous avons donc choisi de suivre une approche analogue à celle de Blanche-Benveniste (2002), qui s'appuie sur les constructions verbales et non sur la phrase pour définir des unités de segmentation. Cependant, si la référence à la construction verbale est très adaptée au monologue narratif ou explicatif, elle doit être étendue pour les corpus de conversation.

Nous partons donc de la **construction,** et par défaut **de la construction verbale** pour définir les unités de segmentation. Nous définissons des « unités maximales de segmentation » (US). Elles sont basées sur des constructions « racines » (root), c'est à dire dont la tête n'a pas de gouverneur. Par défaut, il s'agit de constructions verbales finies, mais aussi nominales, adjectivales ou adverbiales regroupant **un élément tête** ainsi que toutes les séquences qui sont **régies** par lui. Les gouverneurs des constructions racines ne sont eux-mêmes gouvernés par aucune catégorie. Les unités sont donc des **énoncés,** constitués de la séquence « tête + éléments dépendants » étendue aux éléments dits « associés » ou périphériques (en gras) :

à mon avis puisqu'il n'est pas venu *on pourra pas tenir la réunion à Paris //*

je le ferai ***quoi que vous me disiez*** *//*

Dans cette segmentation, un seul symbole est utilisé :

- le symbole // marque la fin d'un énoncé comme dans l'exemple précédent.

Remarque : les énoncés insérés dans d'autres énoncés sous forme de parenthèses ne sont pas marqués par //. Ils sont reconnus lors de l'analyse syntaxique automatique par le lien de dépendance PARENTH (voir section V).

et là on a conscience depuis quelques mois ***enfin c' est ce que je***

ressens *qu' il faudra encore peut-être bien une génération //*

II Conventions pour une segmentation syntaxique reproductible

Pour assurer le caractère reproductible entre annotateurs de cette segmentation sur base syntaxique, des conventions particulières ont été adoptées pour des situations non canoniques. Nous les précisons ci-dessous.

A. Segmentation entre Constructions verbales finies

On distinguera deux situations selon qu'il existe ou non un élément pouvant jouer le rôle de marqueur de frontière d'énoncé.

1. Absence de marqueur

1. 1. Détachements

Sont rattachées à une construction verbale finie associée :

- Les séquences à valeur de lexique détaché avec reprise pronominale:

  *ouais // est-ce que la **la paranoïa** euh **ça** peut euh ça peut être euh une maladie //*

  *et bon c' est pénible parce que ça peut être dangereux hei**n une névrose obsessionnell**e //*

  *c' est c' est tout à fait remarquable et et typique et euh **Djenné**//*

- Les séquences associées sans reprise pronominale (« hanging topic ») :

  *euh donc euh je dirais que **les gens** par rapport à un commerce habituel euh pff déjà c' est pas le même produit //*

- Les séquences détachées à verbe non tensé :

  ex . avec reprise par *tel* d'un sujet infinitif :

  ***le libérer** de toutes les entraves qui freinent son mouvement celles des peuples qui ne peuvent être que populistes celles des syndicats qui ne peuvent être que corporatistes celles des états qui ne peuvent que conduire à l' étatisme celles des nations enfin qui ne peuvent qu' encourager le nationalisme **tel** est le fond de sauce de l' idéologie libérale dominante //*

1.2. Juxtaposition de construction verbales finies

On sur segmente entre deux CV juxtaposées, avec parfois le risque d'une segmentation contre intuitive sémantiquement comme dans :

les bénéfices ça va // ça vient //

On applique la même convention dans le cas de relation sémantique forte (conséquence) comme ci-dessous :

on (n') intervient là-dessus qu' avec du soufre pur euh que l' on poudre //

on poudre les vignes avec ce soufre // ça tue ce champignon //

On segmente aussi sur une base strictement syntaxique dans le cas où plusieurs constructions verbales sont sous la portée de séquences détachées « en facteur commun » :

l' obsessionnel-il va faire des rites // il va aller se laver trente six mille trente six millions de fois les mains pour éviter la souillure dans sa maison dans sa vie //

Certaines séquences verbales tensées non introduites sont cependant regroupées dans une US. C'est le cas des constructions verbales tensées sans marque de dépendance régies par des verbes « présentatifs » (*il y a, ça fait, c'est*). Le critère est que l'ensemble peut être paraphrasé par la seule construction verbale avec laquelle on regroupe la séquence:

j'ai ma sœur elle est malade // = ma sœur est malade
il y a un livre je le trouve pas // = je trouve pas un livre
prenez ma femme elle va jamais au cinéma// = ma femme va jamais au cinéma

C'est le cas également des constructions verbales finies qui sont l'équivalent sémantique et prosodique d'un complément de temps ou de manière :

il était à Paris il y a trois ans // (trois ans auparavant)
il a pris sa retraite ça fait deux ans // (depuis deux ans)
elle s'est mariée elle avait 20 ans // (à 20 ans)
il est arrivé il était en nage // (en nage)

Ou dans les cas de corrélation asyndétiques (*si…alors*)

on trouve un métier ailleurs plus intéressant ben je pars ailleurs pour mieux gagner //

1.3. Discours direct

Seule la première CV du discours direct avec ses dépendants est intégrée à l'US avec le verbe introducteur :

et euh il lui dit euh voilà euh on veut me tuer //on me suit //euh il y a un contrat sur moi //

et il lui dit voilà quand j'arrive dans votre cabinet il y a plus de voix //

1.4. Les pseudo clivées et les « effets deux points »

On ne segmente pas dans le cas d'une pseudo-clivée canonique, où le deuxième terme est considéré comme objet dépendant du premier verbe:

ce qu'il faut dire c'est que les élus en ont assez //

ce qu'il faut dire les élus en ont assez //

On ne segmente pas dans le cas de listes (énumérations) introduites par un terme hyperonyme. Dans de tels exemples, on peut considérer que l'hyperonyme introduit une liste paradigmatique qui pourrait se substituer à lui ou l'inclure :

ça on n' en veut pas des exécutants et puis qui en même temps sont des pompiers inefficaces et impuissants devant les conséquences de cette politique la misère le chômage la délinquance//

On ne segmente pas non plus dans le cas suivant d' « effet deux points »:

// et nous voilà partis donc euh à cinq le Dogon hhh les deux Anglaises mon épouse et moi euh et euh euh et six avec le chauffeur même //

Dans l'exemple suivant, *libérer le marché* est segmenté car on ne peut le considérer comme un objet possible de *existerait.*

mais il ex- il existerait une condition à tout cela évidemment // libérer le marché //

On segmente entre les deux termes des constructions à « spécification lexicale » (pseudo pseudo clivées) qui apparaissent tous deux syntaxiquement complets.

mais auparavant il faut dire une chose // un un élu n' est efficace que s' il est appuyé s' il est en en ligne directe avec ses électeurs //

il faut dire une chose // c'est que les élus en ont assez

1.5. Recours limité au critère prosodique

Les critères morphosyntaxiques de rattachement d'une séquence à un énoncé ne sont pas toujours opératoires. On utilise la prosodie

- dans les cas où un double rattachement d'une séquence est possible :

*il travaille à Paris **dans la confection** // il y a beaucoup d'emplois //*

- dans le cas où plusieurs segmentations sont possibles :

*donc la moindre proposition c' est tout de suite tu vois **est-ce que c' est légal** //*

La prosodie permet de déterminer que la construction verbale « c'est tout de suite » introduit des paroles rapportées analysées comme un complément du verbe.

2 Présence de marqueur

2. 1 Les pauses remplies

Les pauses remplies même associées à une disfluence n'entraînent pas de segmentation :

*ouais est-ce que la la paranoïa euh ça peut **euh** ça peut être euh une maladie //tu vois un type qui **euh** qui est toujours sur les nerfs et qui qui qui est paranoïaque carrément //*

2.2. Les conjonctions

On ne segmente pas :

- Les séquences verbales introduites par une conjonction de subordination :

  donc c' est très long à faire un rosé un vrai rosé parce qu' il faut attendre que la cuve se remplisse //

On considère en effet les prépositions et conjonctions de subordination comme des marques morphologiques permettant de repérer la dépendance à une tête.

La convention est étendue aux relatifs, même dans le cas où on pourrait considérer qu'ils introduisent des relatives « de liaison » :

// surtout quand euh on doit s' occuper d' obsèques je dirais d' enfant de jeune où là ça devient très très pénible //

Remarque : On trouvera cependant des US commençant par une conjonction de subordination dans le cas d'un changement de locuteur (en début de tour de parole).

L1 c'est un mur // L2 hmm // L1 parce qu'en fait l'idée c'était de garder la pente naturelle du terrain //

On segmente :

- Les séquences verbales finies introduites par une conjonction de coordination.

en général c' est comme ça // et euh ce sont donc des blancs des vins blancs // et sur le domaine et en général dans la région on fait des rouges des blancs des rosés et des gris //

On ne segmente pas même dans les cas où la conjonction induit une relation sémantique forte (conséquence), comme dans la dernière US ci-dessous :

donc c' est très long à faire un rosé un vrai rosé parce qu' il faut attendre que la cuve se remplisse / / il faut attendre qu' elle macère // et il faut ouvrir le robinet //et euh quand on est en vendange ben euh on a le feu // donc il faut aller vite // et on peut pas mobiliser beaucoup de cuves comme ça / / donc il faut que ça rentre // et on fait plutôt des gris //

NB . Les séquences régies introduites par des conjonctions de subordination précédées d'une conjonction de coordination avec effet d'épexégèse sont rattachées :

Il le fera mais pas parce que tu le lui as demandé //

B. Enoncés à tête autre que verbe conjugué

Sont considérées comme des énoncés (des US) :

- Les « mots-phrases » ou interjections

On segmente en US les mots isolés qui constituent des énoncés autonomes comme « oui », « ouais », « non », « exactement », « pas du tout », « du tout », lorsqu'ils constituent un tour de parole.

L1 *<oui> //*
L2 *ah oui // oui /*
/

En revanche *voilà, bon, ben, quoi , oui , hein, enfin* sont rattachés à une construction verbale s'ils ne constituent pas un tour de parole distinct. Ils sont reliés au reste de l'énoncé par un lien DM dans l'annotation syntaxique

je suis obsédé ouais ouais // ben il doit être parano hein //

- Les séquences isolées que l'on ne peut rattacher syntaxiquement à un autre énoncé. Le critère est qu'elles sont interprétables indépendamment du contexte comme des « actes de langage » :

en avant //

à toi (de jouer) //
par pitié //

- Les séquences à valeur de commentaire sur une US précédente qui pourraient être introduites par *c'est (un*)

magnifique //
dommage //

il a été attaqué // on lui a tout volé // un vrai désastre //

Remarque : On a distingué les « appositions » à un groupe nominal qui font partie du même énoncé (la même US) comme les coordinations et les énumérations :

il a acheté une voiture une Renault //

des énoncés autonomes formant commentaire qui forment une US :

il a acheté une voiture // un désastre //

- Les structures « binaires » (sans verbe fini comportant au moins deux séquences):

enfin très persécuté très délirant le type //

Dans l'exemple suivant l'adjectif est analysé comme la tête d'une structure binaire intégrée dans une subordonnée:

je dirais que la libre concurrence elle a toujours plus ou moins existé // bon c' est certain que quand c' était l' époque du monopole bon tranquille //

C. Segmentation après disfluences ou abandon de construction

1. Amorces et répétitions

0n ne segmente pas les amorces de constructions ou les répétitions disfluentes (*c'est* dans l'exemple suivant):

donc la moindre proposition c' est tout de suite tu vois est-ce que c' est légal //c' est c' est c' est carrément presque de l' obsession //

Dans ce cas, la segmentation intervient entre deux CV « complètes » *: c'est légal//c'est c'est carrément presque de l'obsession//*. Les amorces de constructions sont intégrées au second énoncé.

2 Constructions abandonnées

A la différence des répétitions simples ou amorces, les bribes ou constructions abandonnées donnent lieu à une segmentation:

alors nous le// euh nous avons a accepté de de les prendre avec nous //

Nous le est une bribe (construction abandonnée) avant le verbe. Mais *a* est une amorce de « accepté » , *de de* est une répétition de la préposition : il n'y a donc pas segmentation.

Dans l'exemple complexe suivant, on a segmenté après la construction abandonnée *qui est toujours* :

// et nous sommes partis // donc euh nous sommes allés à Sanga // donc Sanga qui était euh qui est qui est toujour seuh // en fait nous sommes allés à Sanga pour pour pour qu'//

D. Perturbations entrainées par les tours de parole dans les dialogues

Le parti pris lié au fonctionnement de l'analyseur est le suivant: les paroles de chaque locuteur sont analysées comme des US indépendantes :

L1*il avait honte par rapport aux* ***Marseillais//***
L2 *aux Marseillais //*
L1 ***parce*** *qu'il parlait pas le même provençal qu'eux //*

Donc :

- il y a segmentation en cas d'interruption par un autre locuteur. Même si la séquence interrompue forme une unité syntaxique :

  L2 *je voudrais connaître le nombre de zone de carte Orange //*
  L1 *oui //*
  L2 *pour une personne qui habite le Kremlin Bicêtre rue Pasteur //*

- un énoncé ne peut pas être construit par plusieurs locuteurs même dans les couples question réponse :

  L1 pour aller où //

L2 pour aller euh aux Halles dans dans Paris//

Remarque : Cette contrainte liée au fonctionnement de l'analyseur amène à multiplier les énoncés « fragmentaires », notamment introduits par des prépositions comme dans les énoncés précédents , et de façon particulièrement abusive, dans le cas où la transcription ne représente pas des morphèmes, mais des bruits hors système linguistique *(humm) :*

L1 c'est un mur // L2 humm // L1 parce qu'en fait l'idée c'était de garder la pente naturelle du terrain//

Ces brèves interventions peuvent se superposer au discours d'un locuteur. On a décidé, par convention de les considérer comme des interruptions.

E. Problèmes posés par certaines organisations du discours qui perturbent l'ordre séquentiel syntaxique des énoncés (incises, parenthèses)

1- Les incises

On ne segmente pas comme US les séquences verbales comme « je crois », qui viennent s'insérer dans un discours et qui sont d'un emploi fréquent. Elles sont reliées au reste de l'énoncé par un lien DM dans l'annotation syntaxique.

> *il semblerait que dans la région d' Aigues-Mortes notamment la vigne euh est connue depuis très très très longtemps* ***je crois*** *avant l' an sept cents et qu' à l' arrivée des Arabes pour des raisons de religion la vigne a été obligée de disparaître //*

2- Les parenthèses

Comme on le voit dans les exemples suivants, l'insertion de la parenthèse est d'une nature différente de celle de l'incise. La parenthèse se présente comme une séquence syntaxique libre. Mais comme pour les incises, la parenthèse ne fait pas l'objet de segmentation. Elle est distinguée dans l'annotation syntaxique par un lien PARENT. (Voir guide syntaxique).

> *et là on a conscience depuis quelques mois* ***enfin c' est ce que je ressens qu'*** *il faudra encore peut-être bien une génération //*
>
> *ça m' a ça m' a énormément surpris parce que bon je connais des des régions de l' Allemagne de l' Ouest en particulier la Forêt Noire* ***bon j'y vais de de temps en temps*** *o****ù*** *euh il y a des villages très isolés très perdus mais en même temps extrêmement léchés //*

Voici un exemple avec une incise (paraît-il) dans la parenthèse :

> *et c' est là que la recommandation que m' avait faite euh Omar* ***Omar c' était un descendant du du cheikh Omar*** paraît-*il* ***qui avait euh islamisé le l' Afrique Noire*** *euh m' a servi //*

Remarque : On ne considèrera pas comme une parenthèse des subordonnées insérées avec reprise de la construction :

> *// eh ben Félix il pourrait // même si c' était le meilleur politique du monde il pourrait pas tout faire //*
>
> *par exemple le paranoïaque le l' obsessionnel il va // on va donner l' exemple l' exemple bateau // il ne peut pas prendre le bus //*

(On considère que la partie précédant l'insertion est un énoncé inachevé.)

ni des séquences ajoutées à un terme d'une liste (énumération) :

il est ce qui fut le creuset de dix ans quinze ans ***pour d' autres*** *vingt ans et* ***pour un certain nombre d' entre nous*** *un une peut-être une décennie encore d' amitié //*

F. Enumérations, listes….

Les énumérations (coordinations sans marqueur) et leur extension en listes font partie de la même US que leur gouverneur. Notre annotation n'utilise pas l'équivalent de la virgule à l'écrit.

et en général dans la région on fait des rouges des blancs des rosés et des gris //

Ceci est dû à notre analyse syntaxique de la coordination qui établit une relation directe entre les coordonnés (voir liens PARA et MARK dans le guide d'annotation syntaxique.)