Propositions pour un étiquetage en « séquences fonctionnelles »

Mireille BILGER* & Estelle CAMPIONE**
*Université de Perpignan
**Université de Provence

Résumé

De nombreux chercheurs en Traitement Automatique des Langues ont essayé de mettre au point des stratégies d'analyse partielle (ou *shallow parsing*) des énoncés, faisant apparaître des segments intermédiaires entre le mot et la « phrase », qui sont le plus souvent des suites de mots très courtes, dont le statut linguistique n'est pas très clair et leur caractérisation est généralement faite dans la littérature à l'aide d'exemples plutôt que de définitions précises. Nous adoptons une approche opposée, et nous proposerons une stratégie d'analyse partielle non pas en segments minimaux, mais au contraire en « séquences » maximales qui représentent les constituants fonctionnels rencontrés dans les textes, sans entrer dans le détail de leur composition. En effet, nous faisons l'hypothèse qu'un tel niveau d'analyse, relativement aisé à réaliser manuellement, peut être utile pour l'étiquetage de grands corpus, à partir desquels des analyses plus fines pourront être menées.

1. Introduction

Reconnaissant la difficulté, à la fois théorique et pratique, d'une analyse totale des énoncés, de nombreux chercheurs en Traitement Automatique des Langues ont essayé de mettre au point des stratégies d'analyse partielle (ou *shallow parsing*), faisant apparaître des segments intermédiaires entre le mot et la « phrase ». Le plus souvent, ces segments sont des suites de mots très courtes appelées *chunks* (Abney, 1991) ou

Recherches sur le français parlé, 17, 117-136
©2002, *Université de Provence*

syntagmes non récursifs (Vergne, 1999). Nous empruntons à Vergne (1999) un exemple de découpage :

(La mesure) (de concentration) (de chlorophylle) (est utilisée) (pour estimer) (la biomasse)

Le statut linguistique de ces segments n'est pas très clair, et leur caractérisation est généralement faite dans la littérature à l'aide d'exemples plutôt que de définitions précises, même si des justifications prosodiques (Vergne, 1999) ou psycholinguistiques (Abeillé *et al.*, 2000) sont avancées.

Nous adopterons dans cet article une approche opposée, et nous proposerons une stratégie d'analyse partielle non pas en segments minimaux, mais au contraire en « séquences » maximales qui représentent les constituants fonctionnels rencontrés dans les textes, sans entrer dans le détail de leur composition. Nous faisons l'hypothèse qu'un tel niveau d'analyse, relativement aisé à réaliser manuellement, peut être utile pour l'étiquetage de grands corpus, à partir desquels des analyses plus fines pourront être menées. L'approche utilisée essaie de concilier à la fois une analyse en termes de dépendances et une analyse en termes de constituants, et s'apparente, par certains aspects, aux grammaires en chaînes (*strings grammars*) de Harris (1962) et Salkoff (1973).

2. Description du corpus

Pour tester notre démarche, nous avons appliqué ce découpage sur un corpus de taille réduite, totalisant près de 4000 mots, mais nous avons tenu à ce qu'il soit constitué de trois types de productions ; la répartition entre ces différents types est la suivante :

- un corpus d'extraits de textes écrits tirés de la presse, (*Ça m'intéresse, Sciences et Vie*) totalisant 527 mots ;
- un corpus constitué de textes « fabriqués » [1] destinés à être lus pour une recherche de parole en laboratoire, (*Eurom1*), de 2682 mots ;
- un corpus d'oral spontané (extrait de l'enquête *Corpus de Référence de Français parlé),* de 760 mots.

[1] Il s'agit de la partie française d'un corpus de parole de laboratoire (Eurom1) qui est constitué de passages de cinq phrases liées sémantiquement entre elles et lus par dix locuteurs qui avaient comme consigne de simuler une intonation naturelle.

3. Principes d'analyse

Nous n'avons pas cherché à recenser de manière exhaustive les différentes structures syntaxiques susceptibles d'être rencontrées dans n'importe quel texte. Nous nous sommes appuyées uniquement sur les données qu'offrait notre corpus, et nous nous sommes essentiellement préoccupées des constructions syntaxiques organisées autour d'un verbe.

Les constructions sans verbe rencontrées dans le corpus ont cependant reçu une étiquette qui en signale simplement le type, comme pour les trois exemples suivants :

(1) Const. Adv. (ici les pompiers)

(2) Connecteur (et)
Const. Nom. (moi qui sais pas résister)

(3) Const.Adj. (pareil là-bas)

Nous nous réservons le droit de rendre compte de leur constitution interne et des relations qu'elles peuvent entretenir avec d'autres unités syntaxiques au cours d'une analyse ultérieure.

En revanche, les constructions syntaxiques organisées autour d'un verbe reçoivent une analyse plus fine. Elles vont pouvoir être décomposées en diverses « séquences »[2] qui entretiennent avec cet élément plusieurs types de relation de dépendance. Le calcul de ces relations se fonde sur l'analyse syntaxique développée dans le cadre de l'Approche Pronominale (*cf.* Blanche-Benveniste *et al.*, 1984, 1990).

4. Les séquences

Ce que nous entendons par « séquence » est une unité syntagmatique qui occupe une grande fonction syntaxique, comme celle de Verbe, Sujet, Objet, *etc.* La différence qu'il y a entre ces « séquences » et les habituels « constituants » se situe dans le fait qu'on ne cherche pas à rendre compte des différents éléments qui les constituent. On délimite les séquences en prenant, en fait, « tout ce qui se rapporte à » l'élément fonctionnel (verbe, sujet, objet, *etc.*). On essaie ainsi, en premier lieu, de délimiter les grandes masses qui renvoient à des fonctions. Chaque « séquence » intègre donc des éléments de catégories diverses.

---

[2] Nous avions déjà introduit ce terme de « séquence » dans Bilger (1999).

C N R S ATILF NANCY 2

4.1. La séquence Verbe

Ce que nous désignons par *SéqV* correspond à ce que nous pouvons appeler une construction verbale « minimale », autrement dit, cela correspond à la partie verbale de la construction, dans son extension linéaire. Elle comporte le « bloc verbal » constitué de la forme verbale tensée, des pronoms clitiques « sujet » et « compléments », des particules de négation, ainsi que des participes, des infinitifs, des verbes modaux et de certains adverbes.

Les éléments insérés à l'intérieur de cette séquence ont pour particularité de dépendre de la forme verbale en question et de ne pas pouvoir développer une syntaxe interne. C'est cette double caractéristique qui nous fait intégrer dans cette zone « verbe » les pronoms clitiques « sujet » et « compléments », puisque par définition, ces formes clitiques ne sont pas « développables ». Dans certaines langues (*cf.* Creissel, 1991), cette séquence « Verbe » est d'ailleurs représentée comme une seule unité avec des éléments reliés graphiquement par un trait d'union ou sous forme d'affixes amalgamés sur une base.

Certains adverbes sont liés par des portées réciproques et présentent des dépendances qui leur sont propres, cependant, comme ils sont en définitive branchés sur la forme verbale en question, nous les incluons également dans la séquence « Verbe ».

Voici quelques exemples des différentes extensions que peut prendre la *SéqV* :

- Verbe seul :
  (accompagnait)
- sous forme de liste :
  (construit et dessine)
- pronoms clitiques sujet + objet + verbe :
  (ils les pollinisent)
- verbe + négation :
  (n'est pas)
- verbe + adverbe
  (sont généralement)
- auxiliaire + pp. :
  (ont attaqué)
- faire (laisser) + infinitif :
  (il la fait construire)

- verbe de perception + infinitif :
  (il voyait construire)
- modal + infinitif :
  (devrait s'incurver)
  (il devrait pouvoir les acheter)

La *SéqV* maximale pourrait donc avoir une extension du type :

il ne devrait pas pouvoir les avoir souvent fait polliniser et fertiliser

Les éléments susceptibles de composer la « séquence verbe » semblent pouvoir se calculer facilement. Cependant pour pouvoir la prévoir correctement, il faudrait pouvoir mettre au point la microgrammaire des auxiliaires, des causatifs, des modaux, (*cf.* par exemple, le travail de Chu, 1988). De même, il faudrait connaître la distribution des adverbes qui peuvent venir s'y insérer, et en définir les différentes classes, tant il est vrai qu'à l'intérieur de cette catégorie adverbiale, et notamment en ce qui concerne la négation, il existe un degré d'autonomie non négligeable.

4.2. Les séquences Sujet, Objet et Rection

Nous utilisons des étiquettes telles que *SéqS, SéqO* et *SéqR* pour signaler que la construction verbale se compose, cette fois, de constituants fonctionnels susceptibles de présenter une syntaxe interne. Ces constituants peuvent se réaliser sous une forme lexicale simple ou sous la forme d'une liste. Ils peuvent également comporter des :

- Relatives
- Que-Phrase
- Appositions
- Éléments prépositionnels
- *etc.*

et apparaître sous différents dispositifs, entre autres :

- Extraction en « c'est... que/qui »
- il y a ... qui
- pseudo-clivée

Quel que soit le développement syntagmatique de ces séquences, ces dernières entretiennent obligatoirement une relation de proportionnalité avec une proforme. Cette proforme permet de repérer quelle fonction syntaxique la séquence occupe auprès du verbe constructeur.

Ainsi, l'énoncé suivant :

Cinq voyages vers le Pôle sud et un séjour au Groenland n'ont pas guéri ce glaciologue de la passion qu'il nourrit pour les deux extrémités de la terre, dont il aime étudier les entrailles de glace

est proportionnel à « ils ne l'en ont pas guéri », ce qui permet de dégager quatre séquences du type :

SéqS	(Cinq voyages vers le Pôle sud et un séjour au Groenland)
SéqV	(n'ont pas guéri)
SéqO	(ce glaciologue)
SéqO	(de la passion qu'il nourrit pour les deux extrémités de la terre, dont il aime étudier les entrailles de glace)

Pour l'énoncé du type :

la semaine dernière, mon amie est allée chez le médecin, parce qu'elle avait une angine

le découpage proposé sera le suivant :

SéqR	(la semaine dernière)
SéqS	(mon amie)
SéqV	(est allée)
SéqO	(chez le médecin)
SéqR	(parce qu'elle avait une angine)

Les séquences « Rection » entretiennent une équivalence avec les proformes respectives « quand » et « pourquoi », l'ensemble des séquences *sujet* + *verbe*+ *objet* avec « elle y est allée »

Chercher ces relations de proportionnalité avec les proformes nous permet de rendre compte du schéma syntaxique de la construction mise en cause. Cela nous permet aussi de remettre à plus tard l'analyse de la syntaxe développée à l'intérieur des séquences.

4.2.1. La séquence Sujet (*SéqS*)

Les données du corpus nous ont permis de relever des séquences « Sujet » d'extension différente ; cette séquence se réalise sous une forme simple, sans ajout[3], comme dans :

(chacun) prend parti
(la lumière) se déplace
(Basley) utilise et perfectionne

[3] Nous reprenons ici le terme utilisé, entre autres, par Salkoff (1973).

Elle peut également se réaliser sous la forme d'une liste :

(les petits animaux et les insectes) se sauvent

ou présenter une forme avec ajout(s). Ces ajouts entretiennent alors avec l'élément central plusieurs types de fonctions, entre autres, celles de « complément », de « modificateur » et « d'apposition ». Et, ils vont pouvoir se réaliser dans plusieurs catégories : « nom », introduit ou non par une préposition, « adjectif », « relative », *etc.* En voici quelques exemples relevés dans le corpus :

- ajout de type « complément »

  (les gens du pays)
  (l'accès par la traverse des Sables Jaunes)

- ajout de type « modificateur »
  - sous forme adjectivale

    (des astres très massifs)
  - sous forme de relative

    (l'ambition, qui forme le caractère des gouvernements populaires)
    (le nombre de postes qu'il propose)

- ajout de type « apposition »
  - sous forme nominale

    (Le Pr. Ben Basley, spécialiste des sciences de l'environnement à l'université du Colorado,)
  - sous forme adjectivale :

    (les athéniens, non contents d'avoir délivré leur patrie)
    (des insectes, attirés par le nectar)

  Chaque type d'ajout peut se réaliser sous la forme d'une liste :

  (le jeu de cache-cache du soleil et de la lune)
  (des connexions entre la couronne, le chromosphère et la surface du soleil)

ou bien semblent pouvoir se combiner :

(Amilcar, surnommé Barca, père d'Annibal, d'une famille chère au peuple, soutenu de beaucoup de crédit et d'un grand génie)

et s'imbriquer les uns aux autres de manière récursive :

(la lumière d'une étoile observée au ras du soleil, qui est suffisamment massif pour courber sensiblement l'espace à son voisinage)

La séquence maximale Sujet pourrait donc être du type :

(Le grand Paul de Dijon, dit « le chanceux », frère de Marie et de Marc, originaire d'un quartier cossu de la ville et que sa famille a délaissé)
(a gagné le concours)

la même séquence sous dispositif :

c'est le grand Paul de Dijon, dit « le chanceux ", frère de Marie et de Marc, originaire d'un quartier cossu de la ville et que sa famille a délaissé **qui a gagné** le concours

Les arrangements des différents ajouts sont complexes, et leurs combinatoires difficilement prévisibles.

Ces différents types d'ajouts que nous venons de mentionner sont ceux qui sont traditionnellement recensés pour rendre compte de la possible constitution du syntagme nominal ou « du nom expansé » (cf. Chonin & Delhay, 1998).

Les exemples cités ici n'éliminent pas bien entendu tous les types des SéqS que l'on peut rencontrer dans les textes. La *SéqS* peut se réaliser sous une autre forme que celle d'un syntagme nominal, par exemple, sous la forme d'une « que phrase » ou d'un infinitif :

(que Paul soit si souvent absent) est regrettable
(ne pas le lui avoir donné) était une erreur

4.2.2. La séquence Objet (*SéqO*)

Les compléments qui sont nécessaires au sens et à la construction du verbe, autrement dit les « éléments de valence complément », reçoivent une seule et même étiquette, celle de *SéqO*. Aucune autre information n'est signalée, ni sur la forme que peut prendre le complément en question, ni sur les traits syntactico-sémantiques auxquels il renvoie.

Les « séquences Objet » de catégorie nominale admettent à peu près les mêmes types d'extensions que les *SéqS*, ce qui pourrait donner comme séquence maximale :

(on a photographié)
(le grand Paul de Dijon, dit « le chanceux », frère de Marie et de Marc, vivant dans un quartier cossu de la ville et que sa famille a délaissé)

(on a pensé au)
(grand Paul de Dijon, dit « le chanceux »frère de Marie et de Marc, vivant dans un quartier cossu de la ville et que sa famille a délaissé)

Seul le dispositif aurait sa forme modifiée en « c'est... que » :

c'est le grand Paul de Dijon, dit « le chanceux », frère de Marie et de Marc, vivant dans un quartier cossu de la ville et que sa famille a délaissé qu'on a photographié

Mais bien d'autres types de syntagme pourront être étiquetés comme *SéqO*, entre autres, les « que-phrases », les « si-phrases », les infinitifs, les formes adjectivales, les adverbes. Voici quelques exemples tirés du corpus :

il préfèrerait (que ce soit pas une cuisine végétarienne)
il aimerait savoir (s'il y aura un wagon restaurant)
il n'ose pas (le bouger)
il est désolé (de ne pas avoir pu venir dîner)
il semblait (très proche)
il se comporte (très bien)

4.2.3. La séquence Rection (*SéqR*)

Les éléments qui sont construits par un verbe, mais qui ne le caractérisent pas, sont signalés par l'étiquette *SéqR*. Ceci dit, il n'est pas toujours aisé de statuer sur le côté définitoire ou non que certains éléments peuvent avoir auprès du verbe. Comme cela a été dit dans Blanche-Benveniste (1990), la distinction entre éléments de « valence » et de « rection » est une affaire de lexicographie beaucoup plus que de syntaxe. L'utilisation de dictionnaires de valences tels que celui élaboré dans le cadre du projet Proton[4] (van den Eynde & Mertens, 2001) pourrait donc servir de base à cette distinction.

Dans ce sens, nous n'utilisons pas nécessairement l'étiquette *SéqR*, pour les constituants (prépositionnels ou non), du type :

à Giverny
avec le pépiniériste Joseph Bory
depuis une dizaine d'années
de manière parfaitement correcte

cela dépendra essentiellement du lexique verbal mis en cause. Ainsi pour les verbes, tels que :

habiter (à Giverny)
cohabiter (avec le pépiniériste)
naître (depuis une dizaine d'années)
se comporter (de manière correcte)

ces éléments seront étiquetés comme *SéqO*, dans la mesure où ils semblent être essentiels au sens et à la construction de chacun de ces verbes, alors qu'avec une autre forme verbale, comme par exemple le verbe « peindre », l'ensemble pourra recevoir l'étiquette *SéqR* :

[4] Ce dictionnaire est consultable sur le Web à l'adresse : http://bach.arts.kuleuven.ac.be/PA/protonfr.html (mai 2002).

(peindre des paysages)
(à Giverny)
(avec le pépiniériste)
(depuis une dizaine d'années)
(de manière correcte)

Cette analyse différenciée essaie de s'appuyer au maximum sur des données statistiques, mais en leur absence, elle dépend souvent de la simple intuition que l'on peut avoir de la construction du verbe en question.

Les séquences introduites par un marqueur, tel que :

- parce que
- quand
- comme (comparatif)
- si (d'hypothèse)
- pendant que
- pour que
- *etc.*

si elles sont bien de type rectionnel[5], reçoivent également l'étiquette *SéqR*.

4.3. La séquence Associée (*SéqAss*)

Certains éléments gravitant autour du verbe ont l'apparence de compléments régis par cette forme, ils n'en partagent cependant aucune des propriétés, entre autres : refus de l'extraction en « c'est... que » et aucune équivalence avec une proforme. Nous avons dressé, (*cf.* Blanche-Benveniste et al., 1990), d'une part, une liste des introducteurs de « séquence associée » :

- comme (causal)
- car
- si (de prémisse)
- quant à
- afin que
- puisque
- bien que

[5] Avec certains verbes, les séquences introduites par « quand » ou « comme » sont des compléments valenciels ; elles reçoivent alors l'étiquette « SéqO » :

elle est née quand tu es parti
elle se comporte comme son père le voulait

d'autre part, une liste d'introducteurs qui peuvent, selon le contexte, être des marqueurs de Séquences « associées » ou « rectionnelles », par exemple :

- quand
- parce que[6]
- à mon avis
- selon

La position de la séquence en question, avant ou après la *SéqV*, peut être un critère déterminant, ainsi le « comme (causal) » à valeur d'associé est réalisé avant la *SéqV*. En revanche, le « comme (comparatif) » à valeur rectionnelle se trouve généralement après la *SéqV*.

Il en est de même pour les « séquences » introduites par « si ». Le « si d'hypothèse » rectionnel peut se trouver aussi bien avant qu'après la *séqV*, alors que le « si de prémisse » se trouve généralement avant cette séquence.

Établir ces deux listes, qui ne sont d'ailleurs pas exhaustives, n'épuise pas, bien entendu, tous les cas de *SéqAss* possibles. Bien d'autres formes pourront recevoir aussi cette étiquette :

SéqAss	(heureusement)
SéqV	(il y a pas eu)
SéqO	(de bagarre)

de même que les formes lexicales réalisées sous la forme d'une dislocation, comme dans :

SéqAss	(cette fille)
SéqV	(elle ne parlait pas)
SéqAss	(moi)
SéqV	(je la plains pas)

4.4. La séquence Recteur Faible (*SéqRf*)

On se réserve la possibilité de ne pas traiter automatiquement le premier verbe tensé comme la séquence du verbe constructeur. Si le premier verbe peut être interprété comme un recteur faible, on traite la séquence (y compris le « que ») comme telle. On peut établir une liste de verbes qui pourraient supporter ce statut :

---

[6] Les données du corpus de français parlé nous ont d'ailleurs amenées à vouloir distinguer de manière particulière les SéqAss introduites par « parce que » dans la mesure où elles ont un statut très particulier : elles produisent souvent un effet de « parenthèse » et peuvent avoir un développement relativement important.

- Germain précise que
- la Général a dit textuellement que
- les égyptiens pensaient que
- celle-ci stipule notamment que
- cela signifie que
- ainsi que l'affirme Einstein

on pourrait rajouter à cette liste :

- croire que
- trouver que
- être convaincu que
- *etc.*

Dans le cas où ces formes verbales ont le statut de « recteur faible », c'est-à-dire, dans le cas où on peut les traiter comme des « introducteurs », ces verbes semblent avoir pour caractéristique, entre autres, de ne plus vraiment exercer de sélection sur leur valence.

5. L'étiquette « Insertion »

L'analyse en terme de dépendance ou de non dépendance, par rapport notamment à la *SéqV,* et que nous avons appliquée jusqu'ici, ne permet pas de rendre compte de toutes les composantes rencontrées dans les textes. De fait, certains éléments semblent être « imperméables » à ce type d'analyse ou présentent un intérêt limité à être traités selon ces termes. Ces éléments ont pour particularité de pouvoir s'insérer entre diverses séquences ou à l'intérieur même des séquences, sans en modifier les relations. Nous signalons cette caractéristique par l'étiquette « Insertion » ; nous avons relever trois types d'insertion :

- Insertion Modalité (*InsertMod*)
- Insertion Phatique (*InsertPh*)
- Insertion Parenthèse (*InsertParenth*)

5.1 L'insertion Modalité (*InsertMod*)

Un certain nombre d'adverbes « modaux », comme « peut-être », « sûrement », « sans doute », *etc.*, ont la possibilité d'être suivis de la forme « que » et de précéder aussi bien les séquences réalisées avant la *SéqV* que d'introduire directement cette dernière :

InsertMod	(Peut-être que)
SéqR	(si je lui chante une chanson)
SéqV	(elle se calmera)

SéqR	(si je lui chante une chanson)
InsertMod	(peut-être que)
SéqV	(elle se calmera)

Nous signalons cette possibilité par l'étiquette « *InsertMod* ». Nous utilisons la même étiquette pour le tour interrogatif en « est-ce que » :

InsertMod	(Est-ce que)
SéqR	(si je lui chante une chanson)
SéqV	(elle se calmera)

SéqR	(si je lui chante une chanson)
InsertMod	(est-ce que)
SéqV	(elle se calmera)

5.2 Insertion Phatique (*InsertPh*)

Nous utilisons l'étiquette *InsertPh* pour toute une série de formes qui sont employées comme telles dans les productions et qui peuvent apparemment s'insérer entre n'importe quelle séquence. En voici quelques exemples :

InsertPh	(bon)
SéqAss	(le début de mon voyage)
SéqV	(ça a été)

SéqV	(je suis restée)
SéqO	(trois mois)
InsertPh	(hein)

SéqV	(passez-moi)
InsertPh	(s'il vous plaît)
SéqO	(les réclamations)

5.3. Insertion Parenthèse (*InsertParenth.*)

Nous signalons par cette étiquette, toute construction verbale qui va pouvoir s'insérer entre les différentes séquences déjà citées ou à l'intérieur même de ces séquences, et pour laquelle aucune autre étiquette ne peut être apposée, comme dans l'exemple suivant :

SéqV	(j'ai amené)

SéqO	(des affaires d'hiver des affaires d'été plus euh)
InsertParenth.	(à cette époque j'avais j'étais en maîtrise il me restait le mémoire à faire)
SéqO	(plus euh les livres tout ce qu'il me fallait pour faire mon mémoire)

6. Récapitulatif

Ces différentes séquences permettent de prévoir un enchaînement syntagmatique optimal du type :

SéqAss + SéqRf + InsertMod + SéqS + SéqV + SéqO + SéqR + InsertPh + InsertParenth

L'ordre d'apparition[7] de ces séquences ainsi que leurs points d'insertion sont relativement peu contraints, à l'exception de deux séquences : la séquence « Recteur faible » et la séquence *InsertMod* qui devront nécessairement apparaître avant la séquence « Verbe ».

Nous pouvons proposer un exemple fabriqué qui fait apparaître l'ensemble de ces séquences :

> Évidemment, les collègues ont pensé que si on acceptait d'accompagner l'enseignant, ben peut-être que les jeunes pourraient visiter, je sais pas si visiter est le bon mot, la ville de Paris, parce que c'est comme ça il faut un encadrement

et que l'on peut étiqueter de la manière suivante :

SéqAss	(Évidemment)
SéqRf	(les collègues ont pensé que)
SéqR	(si on acceptait d'accompagner l'enseignant)
InsertPh	(ben)
InsertMod	(peut-être que)
SéqS	(les jeunes)
SéqV	(pourraient visiter)
InsertParenh	(je sais pas si visiter est le bon mot)
SéqO	(la ville de Paris)
SéqAss.PCQ	(parce que c'est comme ça il faut un encadrement)

Il est à noter que nous n'avons pas trouvé une telle suite de séquences dans le corpus que nous avons étudié, ni dans les productions écrites, ni dans les productions orales. Voici quelques extraits étiquetés :

---

[7] Dans un autre cadre, Morel (1997) a également travaillé sur la combinatoire des constituants qui peuvent se trouver en « préambule » de la séquence Verbe.

- Extrait tiré du corpus de presse écrite :

Avec son groupe de recherche du CIRES, le Pr. Ben Balsley, spécialiste des sciences de l'environnement à l'université du Colorado, à Boulder, mène chaque année des campagnes de relevés atmosphériques en divers endroits du globe. Plutôt que les habituels ballons-sondes, Basley utilise et perfectionne depuis une dizaine d'années un système capable d'emporter à plus de 5000 m d'altitude toutes sortes d'instruments de mesure, de maintenir ces capteurs à diverses hauteurs choisies et, surtout, de les récupérer à leur point de départ.

Seul un cerf-volant, ou un ballon captif pour les jours sans vent, peut dresser dans le ciel le câble sur lequel le CIRES fait circuler ses appareillages de mesure.

Science & vie, n°983, août 99, p. 57.

SéqR	(Avec son groupe de recherche du CIRES)
SéqS	(le Pr. Ben Basley, spécialiste des sciences de l'environnement à l'université du Colorado, à Boulder)
SéqV	(mène)
SéqR	(chaque année)
SéqO	(des campagnes de relevés atmosphériques en divers endroits du globe)
SéqAss	(Plutôt que les habituels ballons-sondes)
SéqS	(Basley)
SéqV	(utilise et perfectionne)
SéqR	(depuis une dizaine d'années)
SéqO	(un système capable d'emporter à plus de 5000 m d'altitude toutes sortes d'instruments de mesure, de maintenir ces capteurs à diverses hauteurs choisies et, surtout, de les récupérer à leur point de départ)
SéqS	(Seul un cerf-volant, ou un ballon captif pour les jours sans vent,)
SéqV	(peut dresser)
SéqO	(dans le ciel)
SéqO	(le câble sur lequel le CIRES fait circuler ses appareillages de mesure)

- Extrait tiré du corpus Eurom1 :

j'ai des problèmes avec mon adoucisseur d'eau : le niveau est toujours trop haut, ce qui fait que le trop plein coule sans arrêt. Vous pourriez pas m'envoyer un réparateur mardi matin ? C'est le seul jour de la semaine que j'ai de libre. Si vous êtes d'accord, soyez gentil de me confirmer le rendez-vous par écrit.

SéqV	(j'ai des problèmes)
SéqO	(avec mon adoucisseur d'eau)
SéqS	(le niveau)

SéqV	(est toujours)
SéqO	(trop haut)
SéqRf	(ce qui fait que)
SéqS	(le trop plein)
SéqV	(coule)
SéqO	(sans arrêt)
SéqV	(vous pourriez pas m'envoyer)
SéqO	(un réparateur)
SéqO	(mardi matin)
SéqV	(c'est)
SéqO	(le seul jour de la semaine que j'ai de libre)
SéqAss	(si vous êtes d'accord)
SéqV	(soyez gentil de me confirmer)
SéqO	(le rendez-vous)
SéqO	(par écrit)

- Extrait tiré du corpus de français parlé :

mais au Portugal par contre je suis partie l'hiver - et je ne savais pas pour combien de temps je partais si c'était l'hiver ou plus longtemps ce qui fait que j'ai emmené des affaires d'hiver des affaires euh d'été -plus euh à cette époque j'avais j'étais en maîtrise il me restait le mémoire à faire - plus euh donc euh les livres tout ce qu'il me fallait pour faire mon mémoire là-bas -

Connecteur	(mais)
SéqR	(au Portugal par contre)
SéqV	(je suis partie)
SéqO	(l'hiver)
Connecteur	(et)
SéqV	(je ne savais pas)
SéqO	(pour combien de temps je partais si c'était l'hiver ou plus longtemps)
SéqRf	(ce qui fait que)
SéqV	(j'ai amené)
SéqO	(des affaires d'hiver des affaires euh d'été plus euh)
InsertParenth	(à cette époque j'avais j'étais en maîtrise il me restait le mémoire à faire)
SéqO	(plus euh donc euh les livres tout ce qu'il me fallait pour faire mon mémoire là-bas)
InsertParenth	(mais je m'en rappelle pas là
SéqO	(mais pas beaucoup)
InsertPh	(enfin)
SéqV	(je suis partie donc euh)
InsertPh	(je sais pas)
SéqO	(le vingt-cinq décembre ou le vingt-six un truc comme ça)
Connecteur	(et alors)
Const.Nom	(là dans ma 4L)
SéqV	(j'avais

SéqO	(une 4L fourgonnette)
Connecteur	(et)
SéqV	(je pensais)
InsertPh	(enfin je sais pas)
SéqV	(je pensais)
Connecteur	(mais)
SéqV	(j'avais prévu de dormir)
SéqO	(dans ma 4L)
InsertPh	(enfin bon)
SéqV	(j'avais emmené)
SéqO	(de quoi camper)
InsertPh	(euh enfin comme j'ai dit bon)
séqR	(pour le l'hiver le printemps l'été tout)

7. Distribution

Le tableau 1 donne les fréquences absolues et en pourcentages des différents types de séquences dans les trois corpus. On constate des différences de distributions, qui sont visualisées dans la figure 1. Par exemple, les *SéqS* sont beaucoup plus fréquentes à l'écrit que dans les autres corpus. Afin de juger de la significativité des différences observées, nous avons appliqué un test de χ^2 aux effectifs de chacune des séquences. La colonne p du tableau 1 donne la probabilité d'erreur en affirmant la présence d'un effet significatif : on considère généralement qu'il n'y a pas d'effet significatif au delà d'une valeur de p égale à 0,01 ou 0,05. Les lignes du tableau ont été classées par significativité décroissante.

On voit que les séquences qui discriminent le plus les trois corpus sont :

- *InsertPh* (beaucoup plus représentées à l'oral)
- *SéqS* (beaucoup plus représentées à l'écrit)
- *Connecteur* (beaucoup plus représentées à l'oral)
- *SéqO* (plus représentées dans Eurom)

On observe aussi des différences significatives, mais moins prononcées pour les séquences :

- *SéqR* (plus représentées à l'écrit)
- *SéqAss* (plus représentées à l'écrit)

Par contre la proportion de *SéqV* et *SéqRf* ne diffère pas de façon significative entre les trois corpus.

On notera que nous n'avons pas appliqué un test de χ^2 aux séquences ayant un petit effectif, sur lesquelles nous ne pourrions tirer de conclusions statistiques fiables.

	Ecrit		**Eurom**		**Fref**		**Signif.**	
TypeSéq	*Freq*	*%*	*Freq*	*%*	*Freq*	*%*	*p*	
InsertPh	0	0,0	7	1,0	21	13,7	0,000	oui
SéqS	18	17,6	49	6,8	2	1,3	0,000	oui
Connecteur	5	4,9	39	5,4	22	14,4	0,000	oui
SéqO	25	24,5	252	35,0	36	23,5	0,005	oui
SéqR	17	16,7	64	8,9	10	6,5	0,018	oui
SéqAss	9	8,8	27	3,8	4	2,6	0,032	oui
SéqV	24	23,5	226	31,4	45	29,4	0,258	non
SéqRf	3	2,9	27	3,8	3	2,0	0,519	non
Const.Nom	1	1,0	13	1,8	1	0,7	//	//
InsertMod	0	0,0	10	1,4	1	0,7	//	//
SéqAss.PCQ	0	0,0	0	0,0	4	2,6	//	//
InsertParenth	0	0,0	0	0,0	2	1,3	//	//
Autres	0	0,0	5	0,7	2	1,4	//	//
Total	**102**	**100**	**719**	**100**	**153**	**100**		

Tableau 1. Fréquence des différents types de séquences.

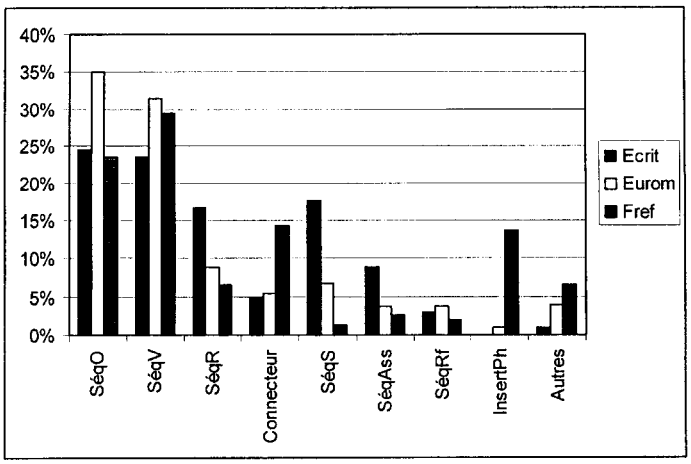

Figure 1. Distribution des séquences en fonction du corpus.

8. Conclusion

Ces propositions d'étiquetage de textes suivis peuvent facilement se prêter à un certain nombre de critiques ; on peut leur reprocher de s'appuyer sur une analyse à la fois trop « compliquée » ou trop « simpliste » ; on peut s'interroger sur l'utilité même de cet étiquetage. Ces propositions doivent avant tout être interprétées comme une première tentative qui nécessite un approfondissement.

Ceci dit, il nous semble qu'en ce qui concerne l'analyse, cet étiquetage en présuppose le moins possible et peut facilement être reproductible. La notion de « séquence fonctionnelle » permet de faire une analyse en grandes masses sans entrer immédiatement dans les détails. Cet aspect est loin d'être négligeable, tant on sait à quel point l'entrée immédiate dans les détails peut être compliquée et propice à des analyses divergentes. Cette procédure minimale de description présente l'avantage de rendre comparables des productions qui pourraient apparaître très différentes les unes des autres de par la simple constitution interne de leurs séquences, de même elle permet de rendre comparables des corpus de langues différentes et d'en faciliter ainsi l'étude contrastive.

Références

Abney, S. (1991). Parsing by chunk. In R. Berwick, S. Abney & C. Tenny (Eds), *Principles-Based Parsing* (pp.257-278). Dordrecht : Kluwer Academic Publishers.

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

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

Bilger, M. (1999). Coordination : analyses syntaxiques et annotations. *Recherches sur le français parlé, 15*, 255-272.

Choi-Jonin, I. & Delhay, C. (1998). *Introduction à la méthodologie en linguistique*. Strasbourg : Presses Universitaires de Strasbourg.

Chu, X-Q. (1989). Examen de la structure de l'enchaînement des verbes modaux. *Recherches sur le français parlé, 9*, 75-86.

Creissels, D (1991). *Description des langues négro-africaines et théorie syntaxique*. Grenoble: Ellug.

Morel, M.-A. (1997). Stratégies intonatives et syntaxe du discours dans l'oral spontané. *Traitement Automatique des langues(T.A.L.), 38-1*, 83-100.

Salkoff, M. (1973). *Une grammaire en chaîne du français. Analyse distributionnelle*. Paris : Dunod.

Vergne, J. (1999). *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur. Analyse syntaxique non combinatoire.* Mémoire d'habilitation à diriger des recherches. Université de Caen, Caen.

Abeillé, A., Clément, L. & Kinyon, A. (2000). Buildling a treebank for French. In *Proceedins of the Second International Conference on Language Resources and Evaluation (LREC 2000),* Athènes.

Harris, Z. (1962). *String analysis of sentence structure.* The Hague : Mouton.

van den Eynde, K. & Mertens, P. (2001). *La syntaxe du verbe, l'approche pronominale et le lexique de valence PROTON.* Preprint n°174, Department of Linguistics, Leuven : K.U. Leuven.
[en ligne : http://bach.arts.kuleuven.ac.be/pmertens/papers/proton.pdf].