Annotation micro- et macrosyntaxique manuelle et automatique de français parlé

Sylvain Kahane[1], José Deulofeu[2], Kim Gerdes[3], Alexis Nasr[2], André Valli[2]

[1] Modyco, Université Paris Nanterre & CNRS
[2] LIF, Aix-Marseille Université & CNRS
[3] LPP, Université Paris 3 Sorbonne Nouvelle & CNRS

1. Le contexte : le projet Orféo

Ce résumé présente les principaux choix d'annotation syntaxique fait dans le cadre du projet Orféo (ANR 2012-2017) visant à fournir un large treebank de français écrit et oral interrogeable en ligne (projet-orfeo.fr). L'annotation est réalisée selon un processus de bootstrapping usuel : un corpus d'amorçage est réalisé par une annotation manuelle, puis un analyseur syntaxique est entraîné sur ce corpus, une nouvelle portion de corpus est analysée automatiquement, puis corrigée manuellement, un nouvel analyseur est entraîné et ainsi de suite. Le corpus comprend plusieurs millions de mots et seule une partie du corpus est corrigée manuellement. Nous utilisons pour la correction manuelle l'Arborator développé par Gerdes (2013) (distribué librement et utilisable en ligne à partir de arborator.ilpga.fr). Plusieurs outils permettant d'entraîner un analyseur en dépendance sont actuellement distribués librement. Nous avons utilisé MATE (Bohnet 2010), ainsi que l'analyseur développé au LIF (Nasr et al. 2011). Le corpus d'amorçage a été développé à partir du treebank Rhapsodie, un corpus de 33 000 mots de français parlé annoté en prosodie et syntaxe distribué librement (Lacheret et al. 2014, projet-rhapsodie.fr) dont l'annotation syntaxique a été entièrement corrigée à la main, à partir d'un pré-annotation automatique réalisée avec un analyseur de l'écrit (de la Clergerie 2005), aucun analyseur pour le français parlé n'étant disponible à l'époque.

Dans ce résumé, nous ne présenterons pas davantage la chaîne de traitement, ni le découpage en unités d'analyse (énoncé, unité illocutoire, phrase, etc.), qui s'avère néanmoins une étape essentielle de l'analyse (voir par ex. Deulofeu et al. 2011, Pietrandrea et al. 2014). Nous allons nous concentrer sur les choix faits pour l'analyse syntaxique en dépendance.

Une remarque préalable essentielle : les choix d'annotation sont toujours un compromis entre diverses exigences (Gerdes & Kahane 2016). Des exigences théoriques : l'annotation doit répondre à un certain nombre de critères fixés par le cadre théorique. Des exigences pratiques liées au processus d'annotation : l'annotation doit être reproductible (accord inter-annotateur), elle doit être la plus simple possible (efficacité, rapidité), et surtout, lorsqu'elle est réalisée en grande partie automatiquement, elle doit pouvoir être propagée sur l'ensemble du corpus en minimisant les erreurs. Enfin des exigences liées à l'utilisateur final : les annotations doivent être facilement requêtables et permettre à l'utilisateur de récupérer les données qu'il souhaite étudier. Nous utilisons l'outil ANNIS (Zeldes et al. 2009) dont le langage de requête permet de décrire des configurations et d'extraire tous les énoncés dont l'arbre syntaxique contient cette configuration.

2. L'annotation syntaxique

De nombreux treebanks en dépendance ont été développés pour un grand nombre de langues (voir par exemple universaldependencies.org). Il existe néanmoins fort peu de treebanks de langues parlés et ceux qui existent, comme le CGN du néerlandais (Oostdijk 2000), ont été construits en appliquant des schémas d'annotation initialement développés pour l'écrit, en gommant notamment certaines spécificités de l'oral comme les disfluences. La principale originalité du projet Orféo est d'être parti de schémas syntaxiques développés dans le cadre de l'analyse du français parlé, notamment l'analyse en grille élaborée autour de Blanche-Benveniste (1990). Les deux traits dominants de ce schéma d'analyse sont la prise en compte de la macrosyntaxe et les listes paradigmatiques.

2.1 Macrosyntaxe

La macrosyntaxe repose sur l'idée qu'une partie des mots d'un énoncé échappe à la rection du prédicat central sans pour autant pouvoir former des énoncés autonomes. On analyse alors un énoncé comme étant constitué d'un noyau, comprenant le prédicat central et pouvant former un énoncé autonome, et de composantes périphériques. Le « tagset » d'Orféo distingue deux principales relations, *periph* pour les composantes périphériques et *dm* pour les marqueurs de discours, qui se distinguent par une plus grande autonomie et une combinatoire moins libre (Kahane & Pietrandrea 2012).

Contrairement à l'annotation de Rhapsodie qui utilisait deux niveaux séparés pour la micro- et la macrosyntaxe (ce qui a permis de mettre en évidence que les contraintes microsyntaxiques peut souvent s'exprimer au-delà des frontières macrosyntaxiques (Deulofeu et al. 2011), Orféo a choisi d'avoir un seul niveau d'annotation encodé à l'aide d'un arbre de dépendance, en privilégiant les relations macrosyntaxiques. Ainsi les compléments détachés à gauche du sujet sont systématiquement annotés *periph*, avec l'idée notamment de simplifier l'annotation pour les humains, comme pour l'analyseur.

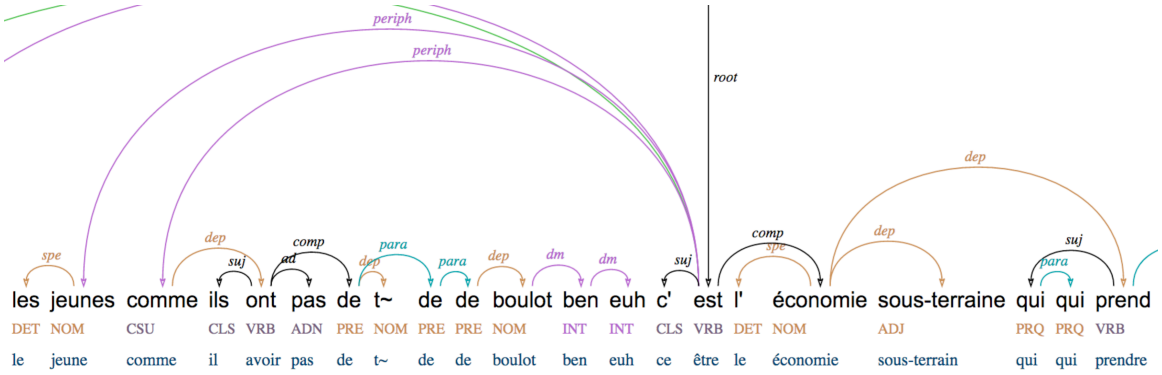

Figure 1. Dépendances macrosyntaxiques *periph* et *dm*

2.2 Listes paradigmatiques

La notion de listes paradigmatiques repose sur la constatation qu'une position syntaxique régie peut être occupée par plusieurs éléments en relation paradigmatique, qu'il s'agisse d'une coordination, d'une reformulation ou même d'une disfluence (Blanche-Benveniste 1990, Gerdes & Kahane 2009, Kahane & Pietrandrea 2012). Dans le cadre d'une syntaxe de dépendance, les listes paradigmatiques peuvent être encodées par l'ajout d'une relation particulière entre les conjoints (relation *para*). Les marqueurs de relation paradigmatique, comme les conjonctions de coordination, sont traités comme dépendant du deuxième conjoint (relation *mark*) (Gerdes & Kahane 2015).

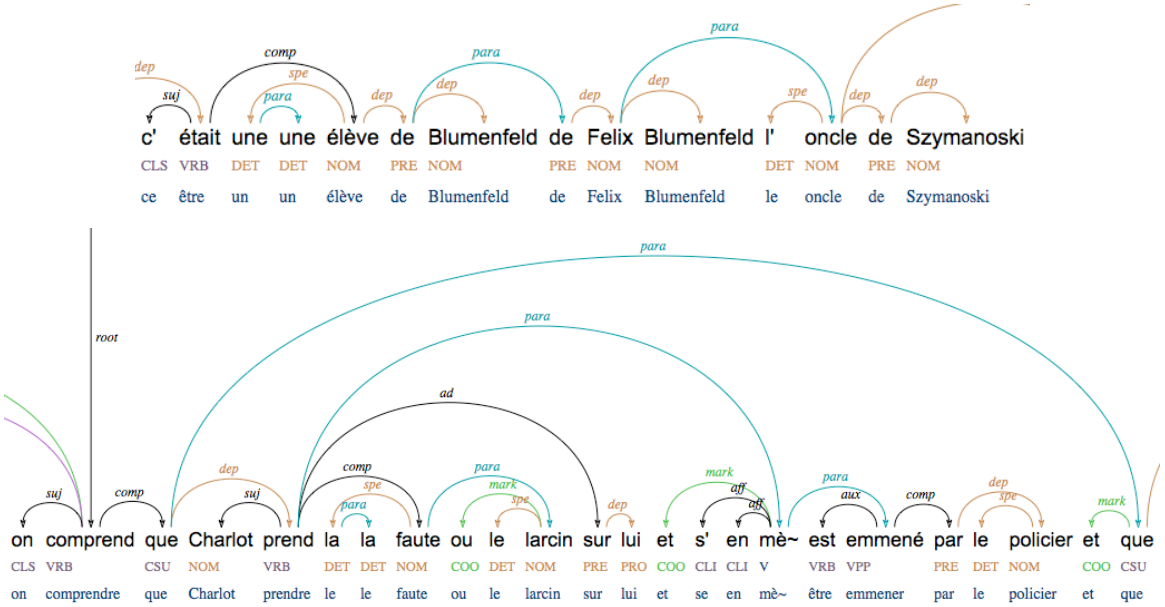

Figure 2. Liens paradigmatiques

3. Conclusion

Les premiers résultats d'entrainement d'un analyseur sur ce schéma d'annotation syntaxique sont extrêmement encourageants. Après entraînement sur seulement 60 000 mots, notre analyseur MATE affiche un f-score de 86,5% pour la reconnaissance du gouverneur d'un mot et 82,6% pour la reconnaissance du gouverneur et de la fonction, ce qui en fait dors et déjà un outil opérationnel pour le traitement automatique du français parlé.[1] Les f-scores pour les relations présentées ici sont : 74,0% pour *periph*, 89,9% pour *dm*, 67,3% pour *para*, 95,1% pour *mark*,

Mots-clés : syntaxe de dépendancemarqueurs de discours, listes paradigmatiques, treebank.

Références

Blanche-Benveniste, C. (1990). Un modèle d'analyse syntaxique « en grilles » pour les productions orales. *Anuario de psicología/The UB Journal of psychology*, (47), 11-28.

Bohnet, B. (2010, August). Very high accuracy and fast dependency parsing is not a contradiction. *ACL*, Uppsala

Deulofeu J., Gerdes K., Kahane S., Pietrandrea P. (2010) Depends on what the French say: Spoken corpus annotation with and beyond syntactic function, *LAW IV*, ACL, Uppsala, 274-281.

K. Gerdes, S. Kahane (2015) Non-constituent coordination and other coordinative constructions as dependency graphs, *Depling*, Uppsala.

Gerdes K., Kahane S. (2016), Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies, *Proceedings of Linguistic Annotation Workshop* (*LAW*), ACL, Berlin.

Kahane, S., Pietrandrea, P. (2012). La typologie des entassements en français. *CMLF,* Lyon, 1809-1828.

Nasr, A., Béchet, F., Rey, J. F., Favre, B., & Le Roux, J. (2011). Macaon: An NLP tool suite for processing word lattices. *ACL-HTL: Systems Demonstrations*, 86-91.

[1] Rappelons que pour l'évaluation, le parseur est entraîné sur 90% du corpus et évalué sur les 10% restant choisis aléatoirement. Le f-score est la moyenne harmonique entre la précision et le rappel.