Annotation syntaxique de français parlé : les choix d'Orféo

Syntactic annotation of Spoken French: Orféo's choices

Sylvain Kahane
Modyco, Université Paris Nanterre & CNRS

Kim Gerdes
LPP, Sorbonne Nouvelle & CNRS

Résumé

Cet article présente les choix d'annotation syntaxique dans le cadre du projet Orféo. Un corpus de français parlé de plus de 200 000 mots a été annoté en syntaxe de dépendance à la main, puis un corpus de 4M de mots a été analysé automatiquement. Les choix d'annotation sont comparés avec ceux du projet Rhapsodie, qui a précédé Orféo, avec Universal Dependencies (UD), qui a démarré un peu après Orféo, et avec Surface-Syntactic UD (SUD), qui est en cours. Orféo se caractérise par une prise en compte de la macrosyntaxe et des phénomènes de listes, ainsi que par un tag set restreint qui a permis une annotation rapide et plus facilement reproductible.

Mots-clés

Corpus annoté, syntaxe de dépendance, syntaxe de l'oral, liste, macrosyntaxe

Abstract

This article presents the syntactic annotation choices for the Orféo project. A corpus of Spoken French of more than 200,000 words was manually annotated in dependency syntax, then a 4M word corpus was automatically parsed. The annotation choices are compared with those of the Rhapsodie project, which preceded Orféo, with Universal Dependencies (UD), which started a little after Orféo, and with Surface-Syntactic UD (SUD), which is in progress. Orféo is characterized by a consideration of macrosyntax and list phenomena, as well as a restricted tag set that allowed a quick and more easily reproducible annotation.

Keywords

Dependency treebank, annotated corpus, spoken language, list, macrosyntax.

1. Le contexte : le projet Orfeo

Notre article présente les principaux choix d'annotation syntaxique faits dans le cadre du projet Orféo (Debaiseux et al., ce volume). La taille du corpus à annoter (10M de mots) et les objectifs du projet (permettre aux utilisateurs, linguistes ou non, de récupérer des exemples d'une construction qui les intéresse) induit nécessairement un compromis entre diverses exigences (voir Gerdes & Kahane 2016

pour une discussion approfondie des critères qui amènent à privilégier un schéma d'annotation plutôt qu'un autre). Des exigences théoriques : l'annotation doit répondre à un certain nombre de propriétés imposées par le cadre théorique. Des exigences pratiques liées au processus d'annotation : l'annotation doit être reproductible (accord inter-annotateurs)[1], 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. Il existe aujourd'hui différents outils permettant d'interroger des treebanks en dépendance, comme l'outil ANNIS (Zeldes et al. 2009) qui était notre choix durant le projet ou GREW (Guillaume et al. 2012, Bonfante et al. 2018) avec lequel nous travaillons maintenant. Ces outils proposent un langage de requête permettant de décrire des configurations (éventuellement complexes) et d'extraire tous les énoncés pour lesquels l'arbre syntaxique contient cette configuration.

Nos exigences théoriques et pratiques nous ont conduits à proposer les choix d'annotation suivants :

- une analyse en dépendance puisqu'elle est économique (on indique pour chaque mot son gouverneur et la nature de la relation qui les unit) et qu'il existe aujourd'hui à la fois de bons outils pour interroger des treebanks en dépendance (voir ci-dessus) et pour apprendre à reproduire de telles analyses automatiquement (Nivre et al. 2007 ; Bohnet 2010 ; Nasr et al., ce volume) ;
- un jeu réduit d'étiquettes syntaxiques, aussi bien du côté des parties du discours que des relations syntaxiques, ce qui permet une annotation manuelle plus efficace et une prise en main plus rapide du schéma d'annotation par les futurs utilisateurs du corpus ;
- une annotation qui prend en compte les principales caractéristiques de la syntaxe de l'oral : l'existence de constituants qui ont une forme d'autonomie syntaxique, comme les constituants détachés ou les marqueurs de discours, l'importance des listes d'éléments occupant une même position, qu'il s'agisse de coordination, de reformulation ou de disfluence ;
- le recours à un lexique de mots grammaticaux comprenant des expressions polylexicales (Deulofeu & Valli, ce volume).

L'annotation a été réalisée selon un processus de bootstrapping usuel : un corpus d'amorçage est annoté manuellement, 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

[1] Pour la grande majorité du corpus, nous avons procédé de manière cumulative : annotation syntaxique automatique, correction, puis vérification par un expert. Seules 300 phrases ont fait l'objet d'une expérience de double annotation à partir des sorties de l'analyseur. L'accord est de 97% pour les parties du discours, 94% pour le nom de la relation, 95% pour le choix du gouverneur (UAS) et 92% pour le choix à la fois de la relation et du gouverneur (LAS). Si l'on regarde uniquement les étiquettes qui ont été modifiées par au moins un des deux annotateurs, les mêmes chiffres passent à 47%, 55%, 60% et 54%.

manuellement. Cette partie corrigée manuellement, qui correspond à ce qu'on appelle traditionnellement le gold, comporte 24 900 phrases et 222 500 lots. Nous avons utilisé pour la correction manuelle l'Arborator développé par Gerdes (2013) (distribué librement sur github.com 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) pour le bootstrapping. Le corpus d'amorçage a été réalisé 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, 2019 ; Kahane et al. 2019) dont l'annotation syntaxique a été entièrement corrigée à la main, à partir d'une 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. Le schéma d'annotation d'Orféo s'est appuyé sur le schéma d'annotation de Rhapsodie (Kahane et al. 2019a,b).

Dans le présent article, nous ne présenterons pas davantage la chaîne de traitement. Nous nous concentrons sur les choix faits pour l'analyse syntaxique en dépendance. La section 2 présentera les relations syntaxiques constituant le cœur de l'analyse syntaxique, à savoir la microsyntaxe. Le cas des listes sera développé dans la section 3. Les relations qui vont au delà de la microsyntaxe et relève de ce que nous appelons, à la suite de Berrendoner (1990) et Blanche-Benveniste (1990), la macrosyntaxe sont présentées dans la section 4. Pour la question des unités minimales de l'analyse, on consultera Deulofeu & Valli (ce volume), et pour celle des unités maxiamles, Nasr et al. (ce volume).

À chaque étape, une comparaison sera effectuée avec le schéma d'annotation de Rhapsodie (lorsqu'il y a une différence notable), avec le schéma d'annotation Universal Dependencies (dorénavant UD, universaldependencies.org, Nivre et al. 2019) dont le développement a été parallèle à celui d'Orféo et le schéma Surface-Syntactic UD (dorénavant SUD, Gerdes et al. 2018, 2019) qui reprend à la fois des caractéristiques d'Orféo et d'UD. On notera que, si de nombreux treebanks en dépendance ont été développés pour un grand nombre de langues (voir le site UD qui regroupe, au jour où nous écrivons cet article, plus de 140 treebanks), 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 la macrosyntaxe et l'analyse en grille des listes, élaborées autour de Blanche-Benveniste (1990).

2. STRUCTURE ET RELATIONS MICROSYNTAXIQUES

Tous les modèles syntaxiques s'accordent sur le fait que les signes linguistiques se combinent pour former des signes linguistiques plus complexes et que c'est ainsi que ce construit la relation entre le signifiant d'un énoncé (un texte vocal ou graphique) et son signifié. L'objectif d'une représentation syntaxique est d'encoder comment les signes linguistiques se combinent, en partant des unités minimales, les

mots et les locutions grammaticales dans notre cas, aux unités maximales. Il existe deux principaux modes de d'encodage des combinaisons : l'analyse en constituants immédiats, qui indique comment une unité se décompose en sous-unités, et l'analyse en dépendance, qui indique par un lien de dépendance la combinaison entre deux unités. La principale différence entre les deux approches concerne la stratification (Kahane 1997, 2018, Kahane & Gerdes, à paraître). Prenons deux exemples. Premier exemple : *Marie aime Pierre* ; une analyse en dépendance indiquera que la forme verbale *aime* se combine à la fois avec son sujet *Marie* et son complément *Pierre*, tandis qu'une analyse en constituant immédiats devra privilégier une des deux combinaisons et dire que cette phrase se décompose en *Marie* + *aime Pierre*, puis *aime Pierre* en *aime* + *Pierre*. Ce qui revient à ordonner les deux combinaisons (du verbe avec son sujet et du verbe avec son objet) et à créer des strates dans la structure. Deuxième exemple : *un livre de syntaxe*. Une analyse en constituants immédiats devra choisir entre une décomposition *un* + *livre de syntaxe* ou *un livre* + *de syntaxe*, tandis qu'une analyse en dépendance dira simplement que *livre* se combine à la fois avec son déterminant *un* et son complément *de syntaxe*.

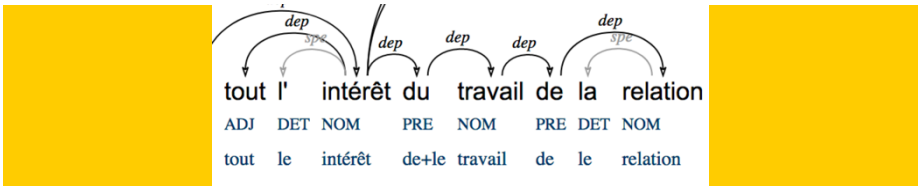
L'autre particularité sur laquelle tous les modèles syntaxiques s'accordent aujourd'hui est que la plupart des combinaisons sont asymétriques. Si l'on considère la combinaison *aime* + *Pierre*, en syntaxe de dépendance, on dira que *aime* gouverne *Pierre* ou inversement que *Pierre* dépend de *aime*, tandis qu'en syntaxe de constituants, on dira que l'unité *aime Pierre* a le constituant immédiat *aime* comme tête syntaxique. La tête d'une unité est, avant tout, l'élément qui contrôle sa distribution (Hudson 1987, Mel'čuk 1988, Kahane & Gerdes, à paraître). Prenons deux exemples. Premier exemple : *à Marie*. Cette unité peut être par exemple complément du verbe PARLER ; elle n'a pas du tout la même distribution que *Marie*, qui peut être sujet ou complément de *aime*. On en déduit que la préposition *à* contrôle la distribution du groupe *à Marie* et gouverne par conséquent *Marie*. Deuxième exemple : *Marie dort*. Ici encore la distribution de *Marie* et *Marie dort* n'ont rien en commun et il est donc clair que *Marie* dépend de *dort*. On peut encore ajouter d'autres critères : par exemple la subordination de la combinaison entre *Marie* et le verbe DORMIR dans *il faut que Marie dorme* affecte la forme du verbe DORMIR, qui est donc l'élément visible par le verbe FALLOIR. Plus généralement, un pur dépendant ne modifie pas la distribution de son gouverneur et est donc invisible au gouverneur de ce gouverneur. En d'autres termes, quel que soit le sujet du verbe DORMIR, la distribution de la combinaison de DORMIR avec son sujet reste la même.

L'application du critère distributionnel distingue notre analyse de celle d'UD qui privilégie les relations entre mots pleins (noms, verbes, adjectifs et adverbes) (voir Osborne & Gerdes (2019) pour une analyse critique de ce choix). Les critères distributionnels tendent eux à attribuer les rôles de têtes aux éléments grammaticaux (on parle alors de têtes fonctionnelles, qu'on oppose à des têtes lexicales). Ainsi dans *parler à Marie*, nous relions *parler* à la préposition *à* qui est la tête (fonctionnelle) de *à Marie*.

À la différence de Rhapsodie, nous avons décidé de ne pas privilégier la tête fonctionnelle dans trois cas (voir la section 3 sur les listes pour le troisième cas). Le premier cas est celui des déterminants : il existe d'assez bons arguments pour traiter le déterminant comme le gouverneur du nom (voir par exemple Hudson 1987 et

Kahane & Gerdes, à paraître). En effet, une unité comme *la syntaxe* n'a pas la même distribution que *syntaxe* : le premier peut être sujet d'un verbe (*la syntaxe m'intéresse*), mais pas le second (**syntaxe m'intéresse*), tandis que le second peut être complément de *parler* (*on a parlé syntaxe toute la nuit*), mais pas le premier (**on a parlé la syntaxe toute la nuit*). Malgré cela, nous avons choisi l'analyse traditionnelle qui fait du nom la tête du « groupe nominal ». En effet, dans certaines positions le nom peut être utilisé avec ou sans déterminant (*le manche de **ce marteau*** vs *un manche de **marteau***), certains noms ont une distribution de modifieurs adverbiaux quel que soit le déterminant qui leur est associé (*il est venu cette **semaine**, il le fera une autre **fois***) et la sémantique du nom contrôle la distribution du groupe (*la famille se réunira* vs **le garçon se réunira*). Le caractère particulier de la construction qui lie le nom au déterminant nous a néanmoins amené à introduire la relation *spe* (pour *specifier*) du nom vers le déterminant (Figure 1).

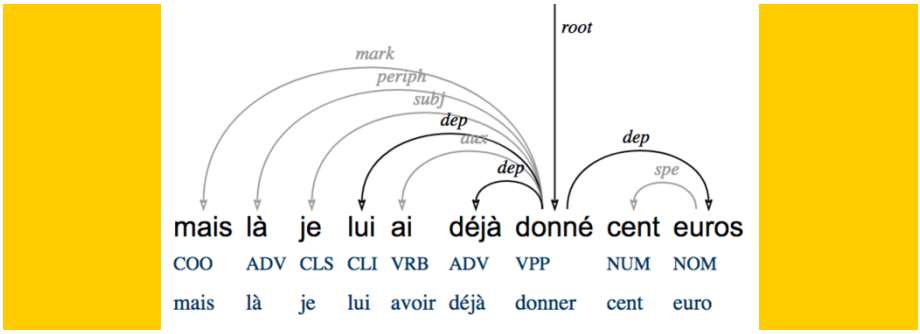
Figure 1. Spécifieurs et dépendants du nom



[TCOF > Lan_reu_mjc_09]

Le deuxième cas est celui des auxiliaires ETRE et AVOIR. On peut vérifier que dans *Marie a dormi*, c'est bien l'auxiliaire *a* qui est la tête. C'est lui qui porte la finitude et qui est affecté lorsque la proposition est subordonnée (*il faut que Marie ait dormi* ; *avoir dormi ne suffit pas*) et c'est bien lui qui impose au verbe *dormir* sa forme de participe passé. Néanmoins, les auxiliaires du français se caractérisent par la montée des clitiques : dans *je lui ai donné ça*, bien que le pronom *lui* dépende de *donné*, celui-ci se cliticise sur l'auxiliaire créant ce qu'on appelle une construction non projective, où le gouverneur de la combinaison *lui* + *donné* se trouve entre eux. La non-projectivité crée de la complexité et les constructions non projectives posent des problèmes à de nombreux analyseurs (Nivre 2006, Nasr et al. 2019). Pour cette raison, nous avons préféré prendre l'auxiliaire comme dépendant du participe par une relation que nous avons nommé *aux* (Figure 2).

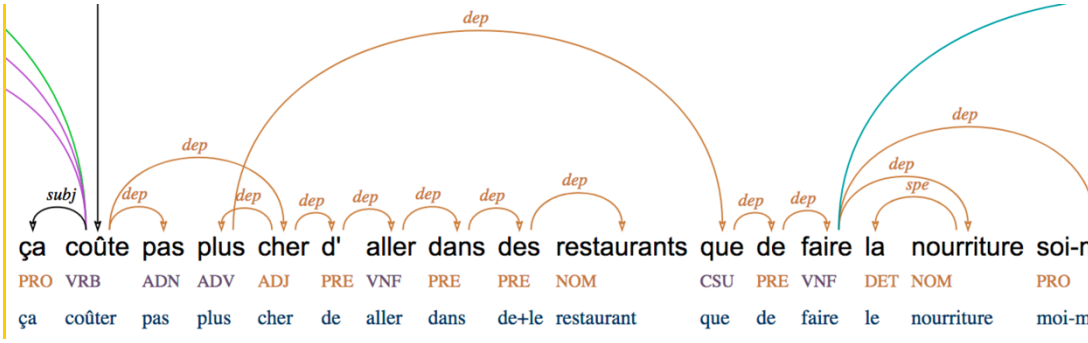
Figure 2. Auxiliaire et clitique



[TCOF > Conv_cai_06]

Certaines constructions du français sont néanmoins analysées avec des constructions non projectives, comme les comparatives (Figure 3).

Figure 3. Construction non projective



[CRFP > PRI-NCY-1]

Le complément *que de faire de la nourriture soi-même* dépend bien du comparatif *plus*, puisque sa suppression entraîne une agrammaticalité (**ça coûte pas cher d'aller dans les restaurants que de faire de a nourriture soi-même*).

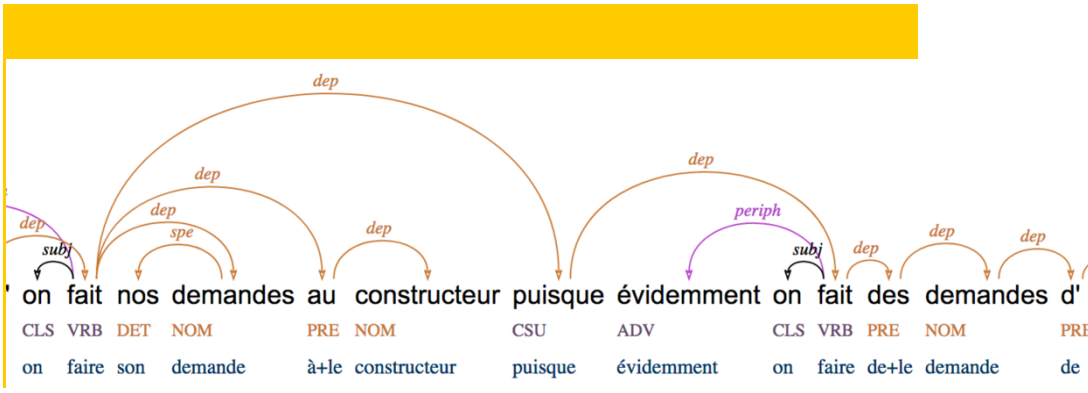
Les autres relations n'ont pas été distinguées et portent toutes l'étiquette *dep* (pour *dependency*), à l'exception des sujets qui dépendent du verbe par la relation *subj* (pour *subject*). Une dernière relation, *disflink* (pour *disfluency link*), a été introduite pour les cas extrêmes où un élément ne rentre pas dans une construction.

Quelques choix d'analyses méritent d'être discutés brièvement.

Les grammaires considèrent en général deux emplois de la forme *des* : un emploi comme déterminant (*j'ai invité des amis*) et un emploi comme combinaison DE + *les* (*on a parlé des impôts*). Nous n'avons pas décomposé les amalgames et *des* est toujours analysé en un seul token (contrairement aux analyses en UD qui proposent de décomposer les amalgames). Mais nous avons été encore plus radicaux en analysant systématiquement *des* (et *du*) comme une préposition (Figure 4). Nous avons fait ce choix, car il est difficile de faire cette distinction de manière automatique sans avoir d'information sur le régime des verbes (INVITER prend un

complément d'objet direct, tandis que PARLER prend un complément d'objet indirect introduit par DE) et les outils actuels font beaucoup d'erreurs.

Figure 4. *au* et *des* comme PRE



[Réunions de travail > OF1_ReunionLocationVoiture]

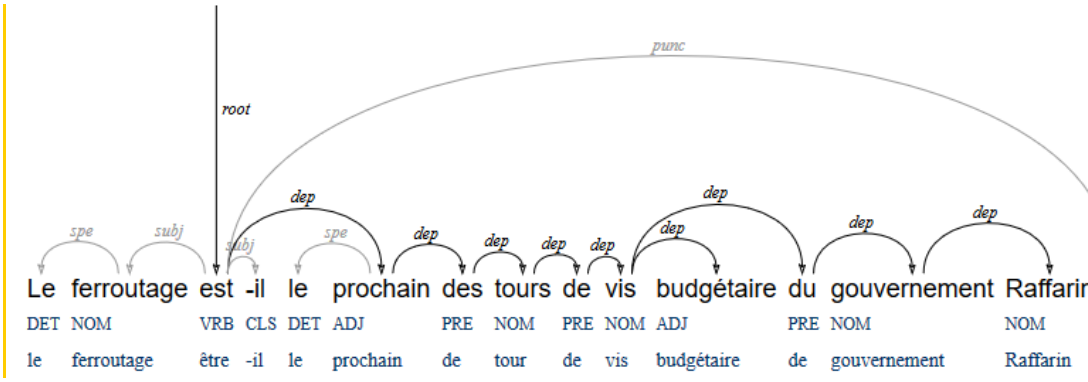
Quelques précisions sur la relation *subj*. Il s'agit du sujet syntaxique. Ainsi dans une phrase comme (2), c'est le pronom impersonnel *il* qui est le sujet du verbe *est*.

1. est-ce qu' il est préférable que l' eau soit acide [TCOF > Aqua_05]

Une analyse plus sémantique comme celle de UD amènerait à considérer le pronom impersonnel *il* comme un explétif et la complétive *que l'eau soit acide* comme le sujet (il s'agit du sujet dit logique ou profond, qu'on appelle encore agent). SUD concilie les deux analyses en indiquant que le pronom personnel est bien le sujet syntaxique tout en étant sémantiquement vide (*subj@expl*) et que la complétive est un complément d'objet à valeur de sujet profond (*comp:obj@agent*).

Dès que le verbe porte un enclitique sujet, celui-ci est déclaré comme sujet. En conséquence, un verbe peut exceptionnellement avoir deux sujets (Figure 5).

Figure 5. Double sujet

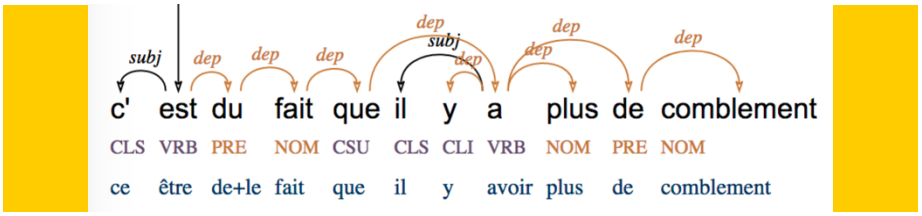


[Chambers-Rostand > L'Humanité]

Cette situation est néanmoins exceptionnelle. En cas de dislocation gauche du sujet, seul le pronom clitique occupant la position microsyntaxique de sujet portera la fonction *subj* (voir Section 4).

De même que les prépositions (PRE) gouvernent le groupe nominal qui les suit, les conjonctions de subordination (CSU) gouvernent la construction verbale qui les suit (Figure 6).

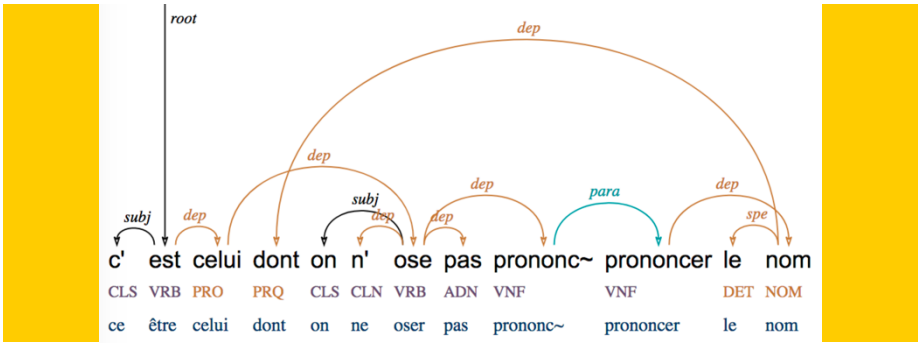
Figure 6. Prépositions et conjonctions de subordination



[TUFS > 27_JD_CP]

Il a été plusieurs fois remarqué que les pronoms relatifs et interrogatifs (PRQ) ont également un rôle de subordonnant (Tesnière 1959 : ch. 246, Hudson 1987, Kahane 2002). Par exemple, si l'on considère la relative *qui dort* (*la fille qui dort est une amie*), on voit que celle-ci ne commute pas avec *Marie dort* et donc que *qui* n'est pas un simple dépendant du verbe *dort*, puisqu'il modifie la distribution de la construction verbale. Une solution possible est d'attribuer deux positions syntaxiques aux PRQ. Pour ne pas compliquer la structure syntaxique (qu'elle reste un arbre de dépendance), nous avons choisi, comme d'autres (et notamment UD), de sacrifier le rôle de subordonnant et de traiter les PRQ comme de simples pronoms. En conséquence, le verbe principal de la relative devient la tête de la relative et dépend du nom antécédent (Figure 7).

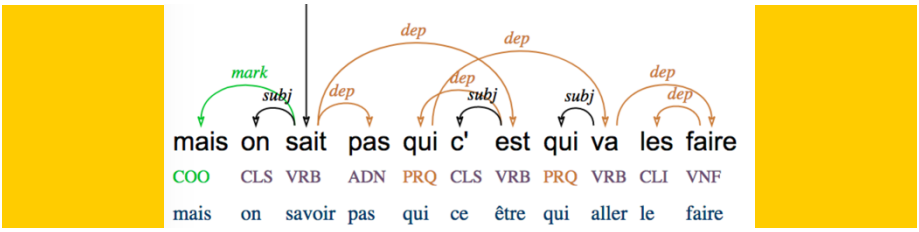
Figure 7. Relative



[Frenchoralnarrative > Kiss_202i-12-13_HELENE_LA_MAGIQUE]

Il en va de même pour les interrogatives indirectes, où le verbe principal de l'interrogative est dépendant du verbe de la principale (Figure 8).

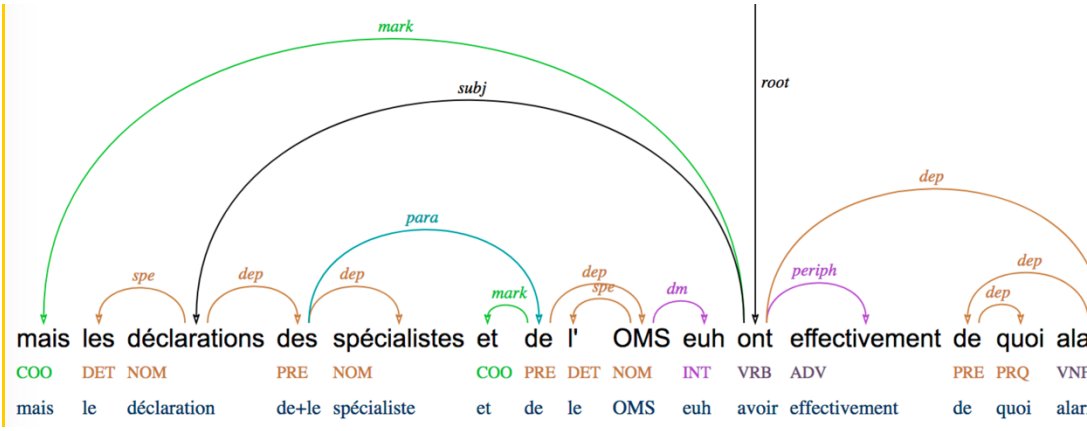
Figure 8. Interrogative indirecte



[TUFS > 27_JD_CP]

Même analyse pour les relatives sans antécédent (Figure 9)

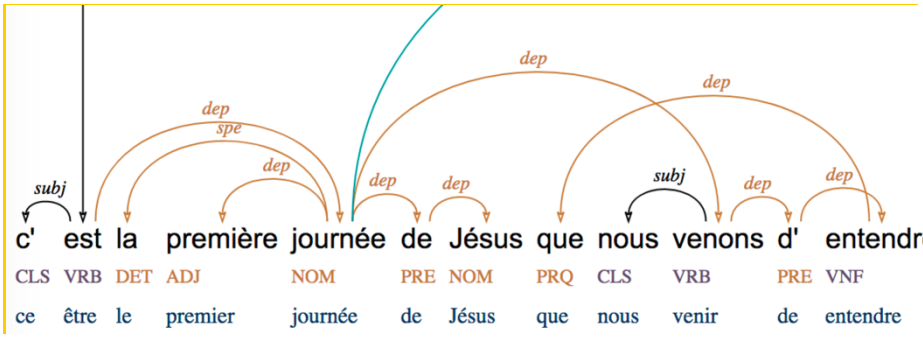
Figure 9. Relative sans antécédent



[Rhapsodie > Rhaps-D2008-RhapsodieBroadcast]

Pour chaque construction clivée qui possède la forme *c'est X qui Y* ou *il y a X qui Y*, la proposition subordonnée dépendra de X. Aucune différence n'est faite entre une construction clivée et une construction "relative présentative". Par conséquent, *c'est un ami qui m'a aidé* et *c'est l'ami qui m'a aidé* seront analysés de façon identique. (La raison en est qu'il ne nous semble pas possible pour un analyseur automatique de discriminer entre les deux situations sans indices prosodiques et pragmatiques.). Cette analyse vaut pour l'objet direct clivé également (Figure 10).

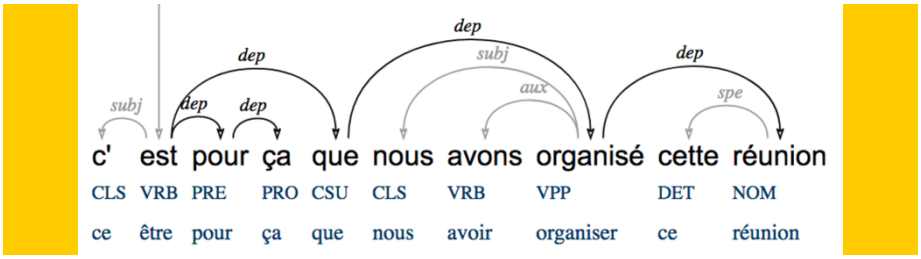
Figure 10. Clivage d'un sujet ou d'un objet



[Rhapsodie > Rhaps-D2003-RhapsodieBroadcast]

Lorsque les propositions clivées présentent un syntagme prépositionnel dans la proposition principale, la proposition subordonnée, qui n'a plus la forme d'une relative standard, est alors analysée comme une complétive et *que* est analysé comme une conjonction de subordination (CSU) (Figure 11).

Figure 11. Clivage d'un groupe prépositionnel

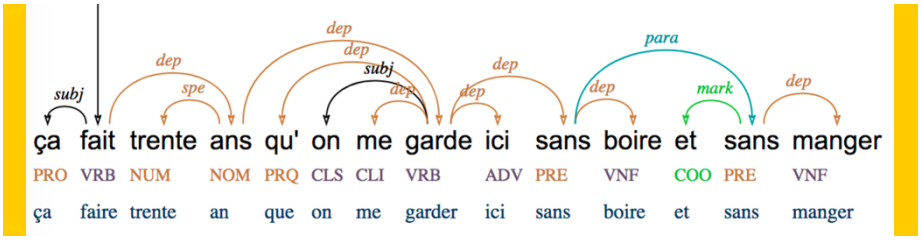


[C-Oral-Rom > fnatps01]

3. LISTES PARADIGMATIQUES

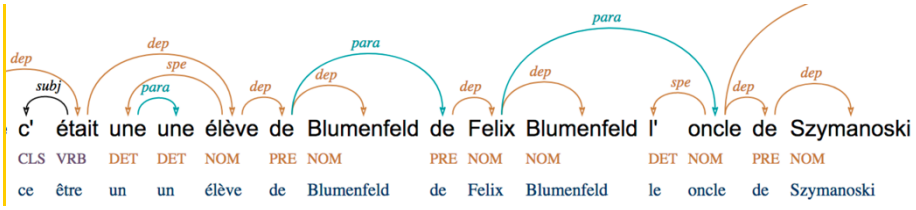
La notion de listes paradigmatiques repose sur la constatation qu'une position syntaxique régie peut être occupée par plusieurs éléments en relation paradigmatique, qu'il s'agisse d'une coordination (Figure 12), d'une disfluence (Figure 7), d'une reformulation ou d'une apposition (Figure 13) (Blanche-Benveniste 1990, Gerdes & Kahane 2009, Kahane & Pietrandrea 2012).

Figure 12. Coordination



[Frenchoralnarrative > Kiss_202i-12-13_HELENE_LA_MAGIQUE]

Figure 13. Reformulation et apposition



[Rhapsodie > Rhaps-D2012-RhapsodieBroadcast]

Les éléments d'une liste paradigmatique, qu'on appelle les conjoints, peuvent occuper seuls la position syntaxique qu'occupe la liste complète. Par exemple, dans l'exemple de la Figure 11, la liste *sans boire et sans manger* a pour conjoints *sans boire* et *sans manger* et chacun d'eux peut commuter avec la liste : *on me garde ici sans boire* ; *on me garde ici sans manger*. Pour cette raison, nous considérons, à la suite de Tesnière (1959) ou Blanche-Benveniste (1990) (voir Kahane (2012) pour une discussion), que les phénomènes de listes sont orthogonaux à la subordination, c'est-à-dire aux relations que nous analysons *dep* ou *subj*. Nous introduisons pour la combinaison des conjoints d'une liste paradigmatique une relation particulière que nous notons *para* (pour liste paradigmatique). La liste est analysée comme une chaîne de conjoints, chaque conjoint dépendant du précédent. Ceci est à contraster avec l'analyse UD qui préfère une analyse en bouquet où tous les conjoints dépendent du premier. Cela revient à décider si chaque nouveau conjoint est combiné avec le précédent conjoint ou avec l'ensemble de la liste qui précède. Une des raisons de préférer l'analyse en chaîne est que plusieurs études ont montré que les langues tendent à minimiser les longueurs des dépendances et donc à lier les mots avec des mots aussi proches que possible (Liu et al. 2017, Futrell et al. 2015, Gildea & Temperley 2010).

Contrairement à Rhapsodie, qui distingue 7 types de listes paradigmatiques (Kahane & Pietrandrea 2012, Kahane et al. 2019b), Orféo n'a qu'un étiquette *para*, la distinction entre coordination et reformulation étant difficile à établir automatiquement en l'absence de marqueurs explicites comme les conjonctions de coordination.[2] UD propose trois relations différentes : *conj* pour les coordinations, *appos* pour les appositions[3] et *reparandum* pour les réparations. Mais comme l'a montré Blanche-Benveniste (1990), les reformulations sont davantage des élaborations par touches successives que des réparations au sens propre.

Les marqueurs de relation paradigmatique, comme les conjonctions de coordination, sont traités comme des dépendants du deuxième conjoint (par une

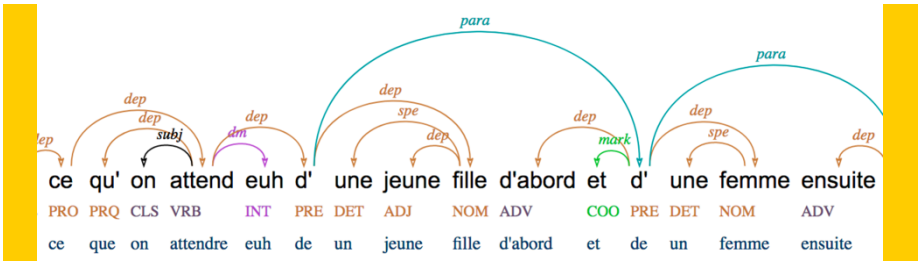
[2] Signalons qu'il s'agit d'une distinction essentiellement sémantique : les conjoints d'une coordination dénote des objets du monde différents, tandis les conjoints d'une reformulation sont différentes dénotations d'un même objet du monde (Kahane & Pietrandrea 2012).

[3] Contrairement aux reformulations, dans les appositions, il s'agit de deux dénotations très différentes, qui donnent deux points de vue sur l'objet du monde, et le conjoint en apposition est en arrière-plan, formant donc une composante périphérique (voir section 5).

relation que nous appelons *mark*) (Gerdes & Kahane 2015). Cette analyse, qui se distingue de l'analyse où la conjonction de coordination est traitée comme la tête du deuxième conjoint (Mel'čuk 1988) et que nous avions appliqué dans Rhapsodie, est justifiée par différentes choses. Premièrement, si la conjonction de coordination a certainement des propriétés de tête, le conjoint reste prépondérant dans la distribution de l'unité qu'ils forment ensemble : ainsi, *et Marie*, *et rouge, et vite* ou *et a mangé* s'ajoutent dans des positions syntaxiques totalement différentes. Deuxièmement, il existe des listes paradigmatiques sans marqueurs réalisés, notamment pour les reformulations, mais aussi dans les coordinations lorsqu'il y a plus de deux conjoints (*Marie, Luc et Pierre*). Troisièmement, nous nous intéressons tout particulièrement aux relations paradigmatiques, c'est-à-dire à la relation qu'entretiennent les éléments qui appartiennent à un même paradigme de commutation.

Un dernier problème mérite d'être mentionné, celui posé par les adverbes dits paradigmatisants (Nølke1983). Les adverbes sont normalement dépendants d'un verbe ou d'un adjectif. Il est néanmoins courant que des adverbes apparaissent dans des entassements paradigmatiques, où ils forment un syntagme avec les conjoints : tel est le cas de *d'abord* et *ensuite* dans l'exemple de la Figure 14. Puisque, dans ce cas, il forme clairement un syntagme avec un conjoint, l'adverbe sera marqué comme un dépendant de la tête du conjoint. Cette analyse pose alors la question de la position syntaxique du même adverbe quand il n'y a plus de liste paradigmatique. Doit-on l'analyser comme dans le cas d'une liste ou bien considérer qu'il dépend du verbe comme le propose l'analyse traditionnelle ? Nous avons généralement opté pour le deuxième cas, mais le problème reste à étudier plus en profondeur de notre point de vue.

Figure 14. Adverbes paradigmatisants



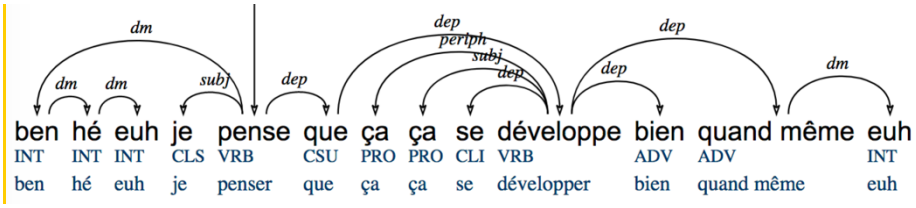
[Rhapsodie > Rhap-D2009-Mertens]

4. 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, Pietrandrea & Kahane 2019) (Figure 15).

Figure 15. *periph* et *dm*



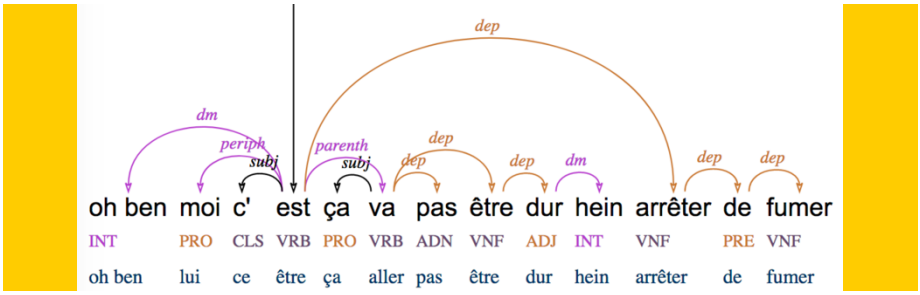
[TUFS > 27_JD_CP]

Nous appelons marqueurs de discours des éléments qui fonctionnent comme des noyaux associés, c'est-à-dire qui portent une forme de force illocutoire propre, qui prédique sur le noyau principal et qui sont généralement fortement lexicalisés et n'acceptent pas de modifieurs. Il peut s'agir d'éléments que les grammaires traditionnelles classent généralement dans les interjections (*ah*, *ouh la la, pff, hein, euh*), mais aussi d'éléments venant des autres catégories (*bon, ben* < bien, *putain*), y compris des constructions verbales (*tu sais, je pense, on dirait*).

Rhapsodie utilisait deux niveaux séparés pour la micro- et la macrosyntaxe, ce qui a permis de mettre en évidence que les contraintes microsyntaxiques peuvent souvent s'exprimer au-delà des frontières macrosyntaxiques (Deulofeu et al. 2010), mais rendait l'analyse des données moins aisées en raison d'encodage différents (la macrosyntaxe était annotée par un balisage du texte et la microsyntaxe par une structure de dépendance). 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 analysés comme des composantes périphériques (et donc annotés *periph*), avec l'idée notamment de simplifier l'annotation pour les humains, comme pour l'analyseur.

Une relation *parenth* a également été introduite pour les parenthétiques (Figure 16). Les parenthétiques se distinguent des marqueurs de discours par leur caractère beaucoup plus libre (on peut ajouter des modifieurs comme dans n'importe quelle construction verbale) et leur complète autonomie illocutoire. Le faible nombre d'occurrences dans notre gold (un peu plus de 200) et l'absence de marqueurs lexicaux fiables n'a pas permis la reconnaissance automatique de cette relation dans le reste du corpus.

Figure 16. Dépendances macrosyntaxiques *parenth, periph* et *dm*



[TCOF > Reso_rich_06]

Le schéma d'annotation UD standard, bien qu'ayant été développé essentiellement à partir de l'analyse de corpus écrits comporte une relation *discourse* similaire à notre relation *dm*. Néanmoins, toutes les composantes périphériques verbales, qu'il s'agisse de marqueurs de discours, comme *je crois,* ou de parenthétiques, sont rattachées par la même relation *parataxis*. Deux relations correspondent à *periph* : la relation *dislocated* pour les compléments détachées non régis et la relation *vocative* pour l'adresse à une des participants d'une discussion. Gerdes & Kahane (2017) proposent divers aménagements du schéma UD pour prendre en compte les distinctions faites par Orféo.

5. Conclusion

Nous avons présenté le schéma d'annotation syntaxique d'Orféo. Un treebank de français parlé de 220 500 mot a été annoté et corrigé manuellement. Un analyseur a été entrainé sur ce treebank gold pour annoter le reste du corpus oral (Nasr et al. 2019). Le schéma d'annotation a été également appliqué à l'écrit, en convertissant les sorties de l'analyseur FRMG (de la Clergerie et al. 2009, de la Clergerie 2013). Les annotations du gold sont disponibles pour l'entrainement d'autres analyseurs et l'ensemble du corpus, ainsi annoté, permet de rechercher des configurations syntaxiques variées.

Remerciements

Nous remercions les collègues qui ont travaillé avec nous à l'élaboration du schéma d'annotation Orféo et tout particulièrement José Deulofeu, Alexis Nasr et André Valli. Nous remercions Jeanne-Marie Debaisieux et Christophe Benzitoun pour l'important travail qu'ils ont accompli dans la mise à disposition des données et dans la gestion du projet. Pour l'annotation syntaxique manuelle, nous remercions Sandra Bellato, Marion Bernard, Sandrine Caddeo, Marine Courtin, Patricia Gori, Marie-Noëlle Roubaud, Frédéric Sabio, Matthieu Stalli. Nous remercions à nouveau Marine Courtin pour les calculs d'accord inter-annotateur.

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.

Blanche-Benveniste, C., Bilger, M., Rouget, C., Van Den Eynde, K., Mertens, P., Willems, D. (1990). *Le français parlé (études grammaticales*). Paris : Editions du CNRS.

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

Bonfante G., Guillaume B., Perrier G. (2018). *Application of Graph Rewriting to Natural Language Processing*. John Wiley & Sons, Incorporated.

De La Clergerie, É., Sagot, B., Nicolas, L., & Guénot, M. L. (2009). FRMG: évolutions d'un analyseur syntaxique TAG du français. In *Journée de l'ATALA sur: Quels analyseurs syntaxiques pour le français?*.

de La Clergerie, É. V. (2013). Improving a symbolic parser through partially supervised learning. In *The 13th International Conference on Parsing Technologies (IWPT)*.

Debaisieux et al (2019), ce volume.

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

Deulofeu J., Valli A. (2019), ce volume.

Futrell, R., Mahowald, K., Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336-10341.

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

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

Gerdes K., Kahane S. (2017) Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral : le cas de la macrosyntaxe, *Actes de la 24e conférence sur le traitement automatique des langues (TALN), Atelier sur les corpus annotés du français* (*ACor4French*).

Gerdes K., Guillaume B., Kahane S., Perrier G. (2018) SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD, *Proceedings of the Universal Dependencies Workshop (UDW)*.

Gerdes K., Guillaume B., Kahane S., Perrier G. (2019) Improving Surface-syntactic Universal Dependencies (SUD): MWEs and deep syntactic features, *Proceedings of the Universal Dependencies Workshop (UDW).*

Gildea D., Temperley D. (2010). Do grammars minimize dependency length?. *Cognitive Science*, *34*(2), 286-310.

Guillaume B., Bonfante G., Masson P., Morey M., Perrier G. (2012). Grew: un outil de réécriture de graphes pour le TAL, *Proceedings of the Joint Conference JEP- TALN-RECITAL*, volume 5: Software Demonstrations.

Hudson R. A. (1987). Zwicky on heads. *Journal of linguistics*, *23*(1), 109-132.

Kahane S. (2002) A propos de la position syntaxique des mots qu-, in P. Le Goffic (éd.), *Interrogation, indéfinition, subordination* [Verbum, XXIV(4)], 399-435.

Kahane S. (2018) Une approche mathématique de la notion de structure syntaxique : raisonner en termes de connexions plutôt que d'unités, *TAL*, 59(1), 13-37.

Kahane, S., Pietrandrea, P. (2012). La typologie des entassements en français. *Actes du Congrès Mondial de Linguistique Fançaise (CMLF),* Lyon, 1809-1828.

Kahane S., Courtin M., Gerdes K. (2018) Multi-word annotation in syntactic treebanks: Propositions for Universal Dependencies, *Proceedings of the 16th international conference on Treebanks and Linguistic Theories (TLT)*, Prague.

Kahane S., Gerdes K. (soumis) *Syntaxe théorique et formelle*.

Kahane S., Gerdes K. , Bawden R. (2019a) Microsyntactic annotation, in Lacheret-Dujour et al. 2019, 49-68.

Kahane S., Pietrandrea P., Gerdes K. (2019b) The annotation of list structures, in Lacheret-Dujour et al. 2019, 69-95.

Lacheret A., Kahane S., Beliao J., Dister A., Gerdes K., Goldman J.-P., Obin N., Pietrandrea P. Tchobanov A. (2014), Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé, *Actes du 4e congrès mondial de linguistique française (CMLF)*, Berlin.

Lacheret-Dujour A., Kahane S., Pietrandrea P. (eds) (2019), Rhapsodie – A Prosodic and Syntactic Treebank for Spoken French, John Benjamins, Amsterdam.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: a new perspective on syntactic patterns in natural languages. *Physics of life reviews*, *21*, 171-193.

Mel'cuk I. (1988) *Dependency syntax: theory and practice*, SUNY press.

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.

Alexis Nasr, Franck Dary, Frédéric Béchet, Benoit Favre (2019) Annotation syntaxique automatique du corpus Orfeo, ce volume.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, *13*(2), 95-135.

Nivre J. et al., 2019, Universal Dependencies 2.4, *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL)*, Faculty of Mathematics and Physics, Charles University, http:// hdl.handle.net/11234/1-2988.

Nølke H. (1983). *Les adverbes paradigmatisants: fonction et analyse* (Vol. 23). Akademisk Forlag.

Osborne T., Gerdes K. (2019). The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics*, *4*(1).

Pietrandrea P., Kahane S., Lacheret A., Sabio F. (2014) The notion of sentence and other discourse units in corpus annotation, in T. Raso, H. Mello, M. Pettorino (eds.), *Spoken Corpora and Linguistic Studies*, John Benjamins, Amsterdam.

Pietrandrea P., Kahane S. (2019) Macrosyntactic annotation, in Lacheret-Dujour et al. 2019, 97-126.

Tesnière L. (1959) *Éléments de syntaxe structurale*, Paris : Klincksieck.