Annotation syntaxique du français parlé : Les choix d'ORFÉO

Sylvain Kahane, Kim Gerdes

Annotation syntaxique du français parlé : les choix d'ORFÉO

Sylvain Kahane, Kim Gerdes

Dans Langages 2020/3 (N° 219), pages 69 à 86
Éditions Armand Colin

ISSN 0458-726X
ISBN 9782200932992
DOI 10.3917/lang.219.0069

Sylvain Kahane
Université Paris Nanterre & Modyco (CNRS UMR 7114)

Kim Gerdes
Université Sorbonne Nouvelle & LPP (CNRS UMR 7018)

Annotation syntaxique du français parlé : les choix d'ORFÉO

1. LE CONTEXTE : LE PROJET ORFÉO

Notre article présente les principaux choix d'annotation syntaxique faits dans le cadre du projet ORFÉO (Debaisieux & Benzitoun 2020 ce volume). La taille du corpus à annoter (9 millions 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 (v. 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 ;
– 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.

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 %.

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, Guillaume & Perrier 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 (cf. *supra*) et pour apprendre à reproduire de telles analyses automatiquement (Nivre *et al*. 2007 ; Bohnet 2010 ; Nasr *et al*. 2020 ce volume) ;
– un jeu réduit d'étiquettes syntaxiques (cf. listes en Annexe), 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 2020 ce volume).

L'annotation a été réalisée selon un processus d'autoamorçage (angl. *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 manuellement. Cette partie corrigée manuellement, qui correspond à ce que l'on appelle traditionnellement ‹le gold›, comporte 183 248 mots. Nous avons utilisé pour la correction manuelle l'ARBORATOR [2] développé par K. Gerdes (2013). 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 ; Lacheret-Dujour, Kahane & Pietrandrea 2019 ; Kahane, Gerdes & Bawden 2019 ; Kahane,

2. Distribué librement sur github.com et utilisable en ligne à partir de arborator.ilpga.fr.

Pietrandrea & Gerdes 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 *et al.* 2009), 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, Gerdes & Bawden 2019 ; Kahane, Pietrandrea & Gerdes 2019).

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 relèvent de ce que nous appelons, à la suite d'A. Berrendonner (1990) et C. Blanche-Benveniste *et al.* (1990), la macrosyntaxe sont présentées dans la section 4. Pour la question des unités minimales de l'analyse, on consultera J. Deulofeu et A. Valli (2020) et pour celle des unités maximales Nasr *et al.* (2020), tous deux dans 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 ; 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 [3]. On notera que, si de nombreux *treebanks* en dépendance ont été développés pour un grand nombre de langues (cf. le site UD qui regroupe, au jour où nous écrivons cet article, des *treebanks* pour une centaine de langues), il existe néanmoins fort peu de *treebanks* de langues parlées et ceux qui existent, comme le CGN du néerlandais (Hoekstra *et al.* 2001), 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 partie 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 C. 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 se construit la relation entre le signifiant d'un énoncé (un texte vocal

3. Les schémas d'annotation UD et SUD sont développés de manière collaborative directement en ligne. Ils peuvent être consultés respectivement sur les sites universaldependencies.org et surfacesyntacticud.github.io. Tous les *treebanks* UD et SUD sont également librement disponibles aux mêmes adresses.

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 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 & Mazziotta 2015 ; Kahane 2018 ; Kahane & Gerdes 2020). 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. Second 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. En effet, 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 2020). La distribution est comprise ici dans un sens purement syntaxique, c.-à-d. la ‹distribution› de l'unité U est l'ensemble des gouverneurs possible de U.

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*. Plus généralement, si la distribution de l'unité AB est différente de la distribution de B alors A contrôle la distribution de AB et A est probablement le gouverneur de B.

(1) $\text{distri(AB)} \neq \text{distri(B)} \Rightarrow \text{A} \rightarrow \text{B}$

Second 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. Plus généralement, si la distribution de l'unité AB est égale à la distribution de A alors B ne contrôle pas la distribution de AB et A est probablement le gouverneur de B.

(2) $\text{distri}(AB) = \text{distri}(A) \Rightarrow A \rightarrow B$

L'application du critère distributionnel distingue notre analyse de celle d'UD qui privilégie les relations entre mots pleins (noms, verbes, adjectifs et adverbes) (v. 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› que l'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 du schéma d'annotation de RHAPSODIE, nous avons décidé de ne pas privilégier la tête fonctionnelle dans trois cas (cf. § 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 (v. p. ex. Hudson 1987 ; Kahane & Gerdes 2020). En effet, une unité comme *la syntaxe* n'a pas la même distribution que *syntaxe* : la première peut être sujet d'un verbe *(la syntaxe m'intéresse)* mais pas la seconde (**syntaxe m'intéresse*) tandis que la seconde peut être complément de *parler (on a parlé syntaxe toute la nuit)* mais pas la première (**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 amenés à introduire la relation *spe* (pour *specifier*) du nom vers le déterminant (Fig. 1) :

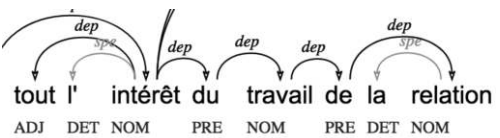

Figure 1 : Spécifieurs et dépendants du nom [ORFÉO_TCOF, Lan_reu_mjc_09]

Le deuxième cas est celui des auxiliaires *être* 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 que l'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 *et al.* 2006 ; Nasr *et al*. 2019). Pour cette raison, nous avons préféré analyser l'auxiliaire comme dépendant du participe par une relation que nous avons nommée *aux* (Fig. 2) :

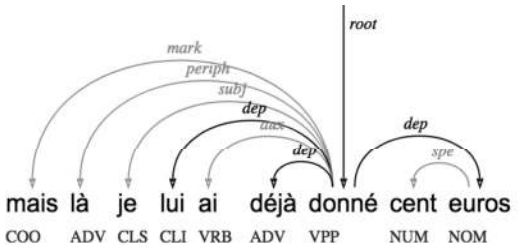

Figure 2 : Auxiliaire et clitique [ORFÉO_TCOF, Conv_cai_06]

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

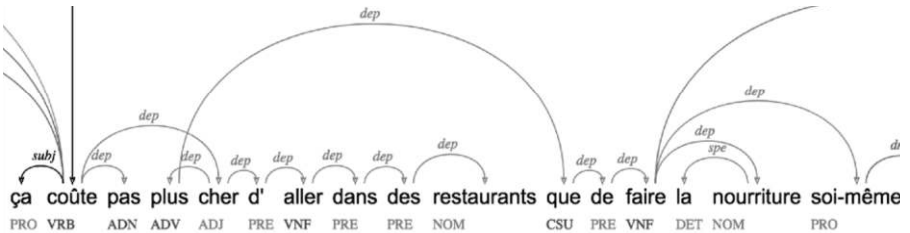

Figure 3 : Construction non projective [ORFÉO_CRFP, PRI-NCY-1]

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 la nourriture soi-même*) alors que la suppression du complément ne pose pas de problème (*ça coûte pas plus cher d'aller dans les restaurants*) ni leur suppression conjointe (*ça coûte pas cher d'aller dans les restaurants*).

En plus des relations *spe* et *aux*, nous avons introduit une relation *subj* pour les sujets. Les autres relations n'ont pas été distinguées et portent toutes l'étiquette *dep* (pour *dependency*). 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)*. Pour l'analyse schématique de *des* et *du* comme prépositions, nous renvoyons à l'article de J. Deulofeu et A. Valli (2020) dans ce volume.

Quelques précisions sur la relation *subj*. Il s'agit du sujet syntaxique. Ainsi, dans une phrase comme (3), c'est le pronom impersonnel *il* qui est le sujet du verbe *est* :

(3) est-ce qu'il est préférable que l'eau soit acide [TCOF, Aqua_05]

Une analyse plus sémantique comme celle d'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›, que l'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 (Fig. 4) :

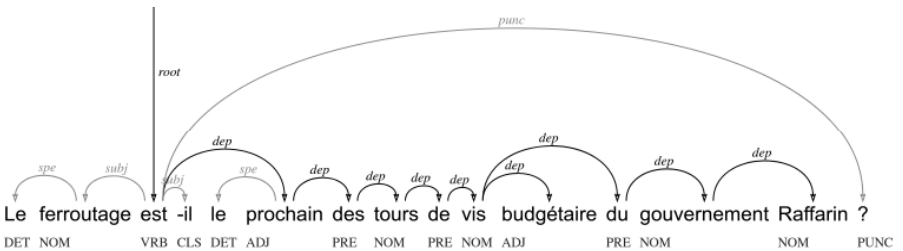

Figure 4 : Double sujet [ORFÉO_CHAMBERS-ROSTAND, L'Humanité]

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* (cf. § 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 (Fig. 5) :

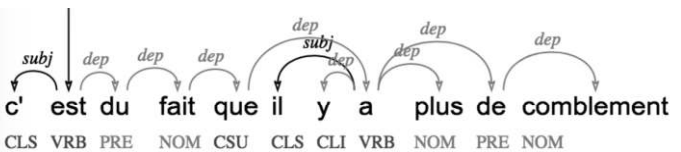

Figure 5 : Prépositions et conjonctions de subordination [ORFÉO_TUFS, 27_JD_CP]

Il a été plusieurs fois remarqué que les pronoms relatifs et interrogatifs (PRQ) ont également un rôle de subordonnant (Tesnière, 1959 : chap. 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 (et pour 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 (Fig. 6) :

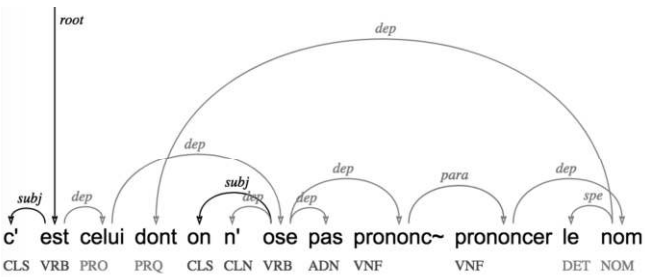

Figure 6 : Relative [ORFÉO_FONC, Kiss_202i-12-13_HELENE_LA_MAGIQUE]

Il en va de même pour les interrogatives indirectes, où le verbe principal de l'interrogative est dépendant du verbe de la principale (Fig. 7) :

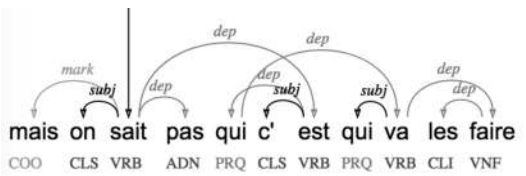

Figure 7 : Interrogative indirecte [ORFÉO_TUFS, 27_JD_CP]

Même analyse pour les relatives sans antécédent (Fig. 8) :

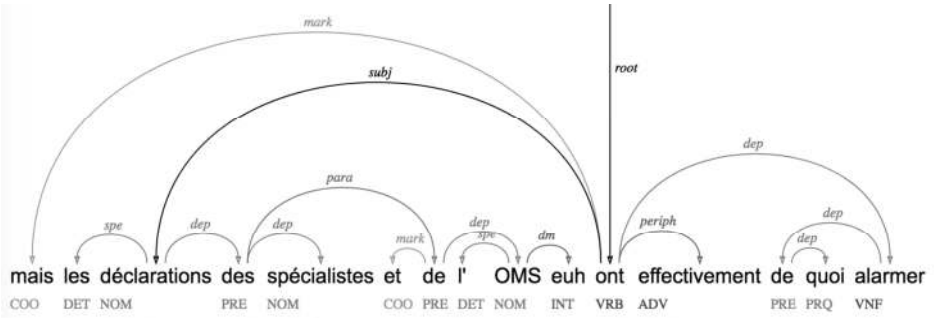

Figure 8 : Relative sans antécédent
[RHAPSODIE, Rhaps-D2008-RhapsodieBroadcast]

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 (Fig. 9) :

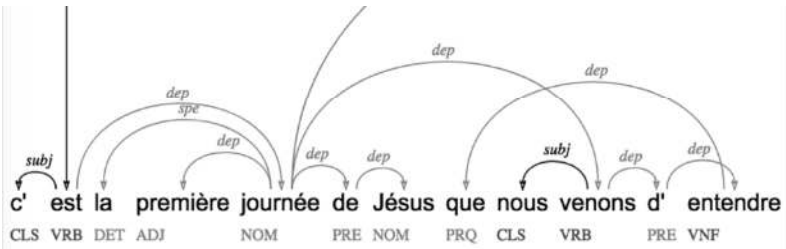

Figure 9 : Clivage d'un sujet ou d'un objet [RHAPSODIE, Rhaps-D2003-RhapsodieBroadcast]

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) (Fig. 10) :

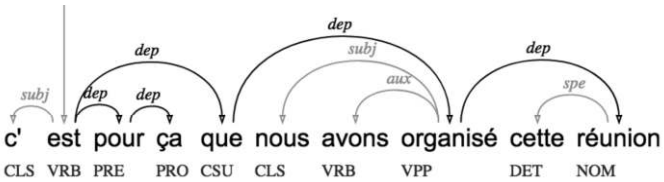

Figure 10 : Clivage d'un groupe prépositionnel [ORFÉO_C-ORAL-ROM, fnatps01]

3. LISTES PARADIGMATIQUES

La notion de *liste paradigmatique* 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 (Fig. 11), d'une disfluence (Fig. 6), d'une reformulation ou d'une apposition (Fig. 12) (Blanche-Benveniste 1990 ; Gerdes & Kahane 2009 ; Kahane & Pietrandrea 2012).

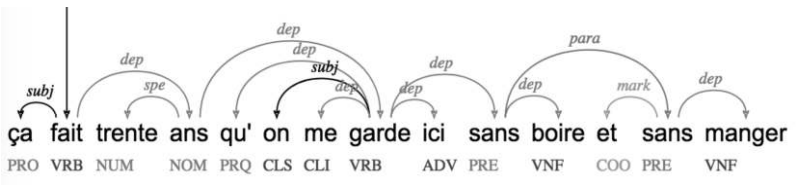

Figure 11 : Coordination
[ORFÉO_FONC, Kiss_202i-12-13_HELENE_LA_MAGIQUE]

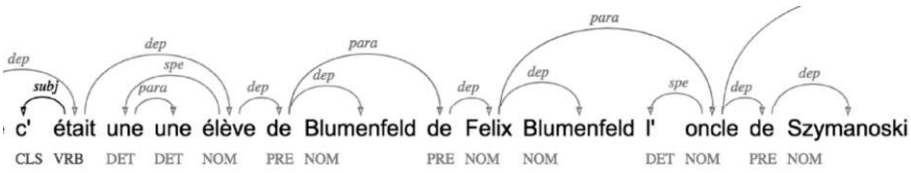

Figure 12 : Reformulation et apposition
[RHAPSODIE, Rhaps-D2012-RhapsodieBroadcast]

Les éléments d'une liste paradigmatique, que l'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 L. Tesnière (1959) ou de C. Blanche-Benveniste (1990) (v. Kahane 2012 pour une discussion), que les phénomènes de listes sont orthogonaux à la subordination, c.-à-d. 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. Cela est à contraster avec l'analyse UD qui préfère une analyse en bouquet où tous les conjoints dépendent du premier, ce qui 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, Xu & Liang 2017 ; Futrell, Mahowald & Gibson 2015 ; Gildea & Temperley 2010).

Contrairement à RHAPSODIE, qui distingue sept types de listes paradigmatiques (Kahane & Pietrandrea 2012 ; Kahane, Pietrandrea & Gerdes 2019), ORFÉO n'a qu'une étiquette *para*, la distinction entre coordination et reformulation étant difficile à établir automatiquement en l'absence de marqueurs explicites comme les conjonctions de coordination [4]. UD propose trois relations différentes : *conj* pour les coordinations, *appos* pour les appositions [5] et *reparandum* pour les réparations. Mais, comme l'a montré C. 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 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 appliquée dans RHAPSODIE, est justifiée par différentes considérations. 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.-à-d. 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› (Nølke 1983). 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 13. 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

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

5. 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 (cf. § 5).

l'analyse traditionnelle ? Nous avons généralement opté pour le second cas mais le problème reste à étudier plus en profondeur de notre point de vue.

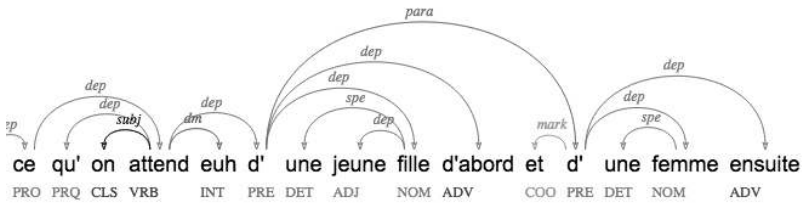

Figure 13 : Adverbes paradigmatisants [RHAPSODIE, Rhap-D2009-Mertens]

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, pouvant former un énoncé autonome, et de composantes périphériques. L'ensemble des étiquettes (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) (Fig. 14) :

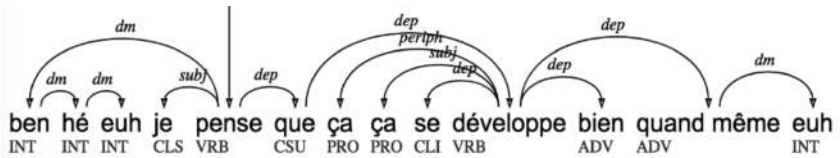

Figure 14 : *periph* et *dm* [ORFÉO_TUFS, 27_JD_CP]

Nous appelons ‹marqueurs de discours› des éléments qui fonctionnent comme des noyaux associés, c.-à-d. qui portent une forme de force illocutoire propre, qui prédique sur le noyau principal, qui sont généralement fortement lexicalisés et qui 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*).

Alors que le schéma d'annotation du projet RHAPSODIE utilise 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 rend l'analyse des données moins aisée en raison d'encodages 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 (Fig. 15). Les parenthétiques sont des unités qui pourraient être des unités illocutoires indépendantes mais elles se trouvent insérées dans une autre unité illocutoire. Elles se distinguent des marqueurs de discours par leur caractère beaucoup plus libre (on peut ajouter des modifieurs comme dans la construction verbale). 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.

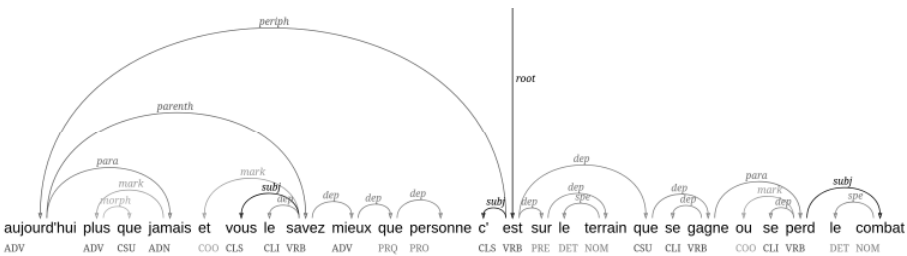

Figure 15 : Dépendances macrosyntaxiques *parenth* et *periph* [RHAPSODIE, Rhap-M2001-C-Prom]

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 détachés non régis et la relation *vocative* pour l'adresse à l'un des participants d'une discussion. K. Gerdes et S. Kahane (2017) proposent divers aménagements du schéma UD pour prendre en compte les distinctions faites par ORFÉO.

5. CONCLUSION

Un corpus de français parlé de 183 248 mots a été annoté en dépendance et corrigé manuellement. Un analyseur a été entraîné sur ce *treebank gold* pour annoter le reste du corpus oral (Nasr *et al.* 2020). 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'entraînement d'autres analyseurs et l'ensemble du corpus, ainsi annoté, permet de rechercher des configurations syntaxiques variées.

Le schéma d'annotation ORFÉO, qui peut s'appliquer aussi bien à des données écrites qu'orales, a été élaboré à partir de données orales en donnant une place importante aux phénomènes paradigmatiques (relation *para*) et macrosyntaxiques (relations *periph*, *dm* et *parenth*). Nous avons fait le choix d'une analyse syntaxique de surface avec un ‹tagset› réduit, qui facilite l'annotation manuelle et automatique. Depuis la fin du projet ORFÉO, le schéma d'annotation UD s'est imposé comme standard, ce qui nous a amenés à développer le schéma d'annotation SUD (SURFACE-SYNTACTIC UD ; Gerdes *et al.* 2018, 2019), qui reprend les caractéristiques majeures du schéma ORFÉO tout en permettant une conversion automatique vers UD.

REMERCIEMENTS

Nous remercions les collègues qui ont travaillé avec nous à l'élaboration du schéma d'annotation ORFÉO et tout particulièrement J. Deulofeu, A. Nasr et A. Valli. Nous remercions J.-M. Debaisieux et C. 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 S. Bellato, M. Bernard, S. Caddeo, M. Courtin, P. Gori, M.-N. Roubaud, F. Sabio, M. Stalli. Nous remercions à nouveau M. Courtin pour les calculs d'accord inter-annotateur.

ANNEXES

Relations de dépendance ORFÉO

dep	Dépendance par défaut
spe	Lien entre nom et déterminant
subj	Sujet
root	Racine
aux	Lien entre verbe lexical et auxiliaire
para	Lien paradigmatique (coordination, reformulation, ...)
mark	Marqueur d'un lien paradigmatique (entre conjoint et conjonction de coordination)
dm	Marqueur de discours
periph	Composante périphérique
parenth	Parenthétique
disflink	Lien pour les segments abandonnés non paradigmatiques
punc	Ponctuation

Parties du discours ORFÉO

NOM	Nom
DET	Déterminant
ADJ	Adjectif
NUM	Numéral
PRE	Préposition
PRO	Pronom (par défaut)
PRQ	Pronom relatif et interrogatif
VRB	Verbe fini
VNF	Verbe à l'infinitif
VPP	Verbe au participe passé
VPR	Verbe au participe présent
ADN	Adverbe négatif
ADV	Adverbe (par défaut)
CLI	Pronom clitique (par défaut)
CLN	Clitique négatif (*ne*)
CLS	Clitique sujet
CSU	Conjonction de subordination
COO	Conjonction de coordination
INT	Interjection
X	Inanalysable (amorce, mot étranger)
PUNC	Ponctuation

Références

[ORFÉO] *Corpus d'Étude pour le Français Contemporain*, ATILF, LIF, Loria, CLLE-ERSS, ICAR, LaTTiCe. [https://www.ortolang.fr/market/corpora/cefc-orfeo]

BERRENDONNER A. (1990), « Pour une macro-syntaxe », *Travaux de linguistique* 21, 25-36.

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. *et alii* (1990), *Le français parlé : études grammaticales*, Paris, Éditions du CNRS.

BOHNET B. (2010), "Very high accuracy and fast dependency parsing is not a contradiction", *Proceedings of the 23rd International Conference on Computational Linguistics – COLING'10*, Stroudsburg (PA), Association for Computational Linguistics, 89-97.

BONFANTE G., GUILLAUME B. & PERRIER G. (2018), *Application of Graph Rewriting to Natural Language Processing*, London/Hoboken (NJ), ISTE/Wiley.

DE LA CLERGERIE É. (2013), "Improving a symbolic parser through partially supervised learning", in H. Bunt, K. Sima'an & L. Huang (eds.), *Proceedings of the 13th International Conference on Parsing Technologies – IWPT 2013* (Nara, Japan), Stroudsburg (PA), Association for Computational Linguistics, 54-72.

DE LA CLERGERIE É. *et alii* (2009), « FRMG : évolutions d'un analyseur syntaxique TAG du français », *Actes de la Journée « Quels analyseurs syntaxiques pour le français ? »* (Paris, France), Orsay, ATALA. [en ligne]

DEBAISIEUX J.-M. & BENZITOUN C. (2020), « Présentation », *Langages* 219. (ce volume)

DEULOFEU J. & VALLI A. (2020), « Lexique et classement en parties du discours dans ORFÉO », *Langages* 219. (ce volume)

DEULOFEU J. *et alii* (2010), "Depends on what the French say: Spoken corpus annotation with and beyond syntactic function", *Proceedings of the Fourth Linguistic Annotation Workshop – LAW IV*, Stroudsburg (PA), Association for Computational Linguistics, 274-281.

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. (2013), "Collaborative dependency annotation", in E. Hajičová, K. Gerdes & L. Wanner (eds.), *Proceedings of the Second International Conference on Dependency Linguistics – DepLing 2013* (Prague, Czech Republic), Prague, Charles University in Prague & Matfyzpress, 88-97.

GERDES K. & KAHANE S. (2009), "Speaking in piles: Paradigmatic annotation of a French spoken corpus", in M. Mahlberg, V. González-Díaz & C. Smith (eds.), *Proceedings of the Fifth Corpus Linguistics Conference – CL2009* (Liverpool, UK), article #309. [en ligne]

GERDES K. & KAHANE S. (2015), "Non-constituent coordination and other coordinative constructions as dependency graphs", in J. Nivre & E. Hajičová (eds.), *Proceedings of the Third International Conference on Dependency Linguistics – DepLing 2015* (Uppsala, Sweden), Stroudsburg (PA), Association for Computational Linguistics, 101-110.

GERDES K. & KAHANE S. (2016), "Dependency annotation choices: Assessing theoretical and practical issues of universal dependencies", in A. Friedrich & K. Tomanek (eds.), *Proceedings of the 10th Linguistic Annotation Workshop – LAW-X 2016*, Stroudsburg (PA), Association for Computational Linguistics, 131-140.

GERDES K. & KAHANE S. (2017), « Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral : le cas de la macrosyntaxe », dans L. Danlos *et alii* (éds), *Actes de la 24e Conférence sur le Traitement Automatique des Langues Naturelles – TALN, Atelier sur les corpus annotés du français – ACor4French 2017* (Orléans, France), Orsay, ATALA, 1-9.

GERDES K. *et alii* (2018), "SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD", *Proceedings of the Second Workshop on Universal Dependencies – UDW 2018* (Brussels, Belgium), Stroudsburg (PA), Association for Computational Linguistics, 66-74.

GERDES K. *et alii* (2019), "Improving Surface-syntactic Universal Dependencies (SUD): MWEs and deep syntactic features", in M. Candito *et alii* (eds.), *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories – TLT, SyntaxFest 2019* (Paris, France), Stroudsburg (PA), Association for Computational Linguistics, 126-132.

GILDEA D. & TEMPERLEY D. (2010), "Do grammars minimize dependency length?", *Cognitive Science* 34 (2), 286-310.

GUILLAUME B. *et alii* (2012), « Grew : un outil de réécriture de graphes pour le TAL », dans L. Besacier, H. Blanchon & G. Sérasset (éds), *Actes de la conférence conjointe JEP-TALN-RECITAL 2012* (Grenoble, France), vol. 5, 1-2. [en ligne]

HOEKSTRA H. *et alii* (2001), "Syntactic annotation for the spoken Dutch corpus project (CGN)", in W. Daelemans *et alii* (eds.), *Computational Linguistics in the Netherlands 2000*, Amsterdam, Brill | Rodopi, 73-87.

HUDSON R. A. (1987), "Zwicky on heads", *Journal of Linguistics* 23 (1), 109-132.

KAHANE S. (2002), « À propos de la position syntaxique des mots *qu-* », *Verbum* XXIV (4), 399-435.

KAHANE S. (2012), « De l'analyse en grille à la modélisation des entassements », dans S. Caddéo *et alii* (éds), *Penser les langues avec Claire Blanche-Benveniste*, Aix-en-Provence, Presses Universitaires de Provence, 101-116.

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. & GERDES K. (2020), *Syntaxe théorique et formelle*, Berlin, Language Science Press.

KAHANE S. & MAZZIOTTA N (2015), "Syntactic polygraphs: A formalism extending both constituency and dependency", in M. Kuhlmann, M. Kanazawa & G. M. Kobele (eds.), *Proceedings of the 14th Meeting on the Mathematics of Langage – MoL 2015* (Chicago, USA), Stroudsburg (PA), Association for Computational Linguistics, 152-164.

KAHANE S. & PIETRANDREA P. (2012), « La typologie des entassements en français », dans F. Neveu *et alii* (éds), *Actes du 3e Congrès Mondial de Linguistique Fançaise – CMLF 2012* (Lyon, France), Les Ulis, EDP Sciences, 1809-1828.

KAHANE S., GERDES K. & BAWDEN R. (2019), "Microsyntactic annotation", in A. Lacheret-Dujour, S. Kahane & P. Pietrandrea (eds.), *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*, Amsterdam, John Benjamins, 49-68.

KAHANE S., PIETRANDREA P. & GERDES K. (2019), "The annotation of list structures", in A. Lacheret-Dujour, S. Kahane & P. Pietrandrea (eds.), *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*, Amsterdam, John Benjamins, 69-95.

LACHERET-DUJOUR A., KAHANE S. & PIETRANDREA P. (eds.) (2019), *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*, Amsterdam, John Benjamins.

LACHERET A. *et alii* (2014), Rhapsodie : un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé, dans F. Neveu *et alii* (éds), *Actes du 4e Congrès Mondial de Linguistique Française – CMLF 2014* (Berlin, Allemagne), Les Ulis, EDP Sciences, 2675-2689.

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'ČUK I. (1988), *Dependency Syntax: Theory and Practice*, Albany (NY), SUNY press.

NASR A. *et alii* (2011), "Macaon: An NLP tool suite for processing word lattices", *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – ACL-HLT 2011, Proceedings of System Demonstrations* (Portland, Oregon, USA), Stroudsburg (PA), Association for Computational Linguistics, 86-91.

NASR A. *et alii* (2020), « Annotation syntaxique automatique de la partie orale du CÉFC », *Langages* 219. (ce volume)

NIVRE J. *et alii* (2006), "Labeled pseudo-projective dependency parsing with support vector machines", in L. Màrquez & D. Klein (eds.), *Proceedings of the Tenth Conference on Computational Natural Language Learning – CoNLL-X* (New York City, USA), Stroudsburg (PA), Association for Computational Linguistics, 221-225.

NIVRE J. *et alii* (2007), "MaltParser: A language-independent system for data-driven dependency parsing", *Natural Language Engineering* 13 (2), 95-135.

NIVRE J. *et alii* (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. [en ligne]

NØLKE H. (1983), *Les adverbes paradigmatisants : fonction et analyse*, Copenhague, 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), 17. [en ligne]

PIETRANDREA P. & KAHANE S. (2019), "Macrosyntactic annotation", in A. Lacheret-Dujour, S. Kahane & P. Pietrandrea (eds.), *Rhapsodie: A Prosodic and Syntactic Treebank for Spoken French*, Amsterdam, John Benjamins, 97-126.

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

ZELDES A. *et alii* (2009), "ANNIS: A search tool for multi-layer annotated corpora", in M. Mahlberg, V. González-Díaz & C. Smith (eds.), *Proceedings of the Fifth Corpus Linguistics Conference – CL2009* (Liverpool, UK), article #358. [en ligne]