Annoter la parole spontanée en arbres de constituants pour les besoins de l'analyse temporelle : résultats et comparaison français parlé / français écrit

Ilaine Wang, Jean-Yves Antoine, Lotfi Abouda, Jakub Waszczuk, Aurore Pelletier, Anaïs Halftermeyer

▶ **To cite this version:**

Ilaine Wang, Jean-Yves Antoine, Lotfi Abouda, Jakub Waszczuk, Aurore Pelletier, et al.. Annoter la parole spontanée en arbres de constituants pour les besoins de l'analyse temporelle : résultats et comparaison français parlé / français écrit. Congrès Mondial de Linguistique Française, Jul 2020, Montpellier, France. hal-02523135

HAL Id: hal-02523135
https://hal.science/hal-02523135

Submitted on 28 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Annoter la parole spontanée en arbres de constituants pour les besoins de l'analyse temporelle : résultats et comparaison français parlé / français écrit

Ilaine Wang[1,2], *Jean-Yves* Antoine[1], *Lotfi* Abouda[3], *Jakub* Waszczuk[4], *Aurore* Pelletier[1] et *Anaïs* Halftermeyer[2]

[1]LIFAT, Université de Tours, 41000 Blois, France
[2]LIFO, Université d'Orléans, 45000 Orléans, France
[3]LLL, Université d'Orléans, 45000 Orléans, France
[4]Institut für Sprache und Information, Heinrich-Heine-Universität, Düsseldorf, Allemagne

Résumé. Cet article présente les principaux résultats de la partie syntaxique du projet Temporal@ODIL, une initiative visant la construction d'un corpus de français parlé spontané annoté en temporalité. Nous présentons ici ODIL_Syntax, corpus arboré en constituants sur lequel s'appuie l'annotation temporelle et qui est diffusé librement sous licence Creative Commons. ODIL_Syntax a été créé à l'aide de `Contemplata`, une plateforme Web d'annotation développée spécifiquement dans le cadre du projet, diffusée elle aussi librement et qui présente l'intérêt de permettre une annotation semi-automatique utilisant un analyseur syntaxique. L'article décrit la procédure d'annotation avec cet outil, nos choix d'annotation ainsi que le corpus produit, en s'intéressant en particulier à une comparaison avec le corpus équivalent FTB (*French Treebank*) développé pour l'écrit.

Abstract. Constituency annotation of spontaneous speech for temporal analysis needs: results and comparison between spoken and written French. This paper presents the main results drawn from the syntactic part of Temporal@ODIL, a project whose objective is the construction of a temporally annotated corpus of spontaneous speech for French. We describe ODIL_Syntax, a freely distributed constituency treebank on which our temporal annotation is grounded. The syntactic annotation was performed on Contemplata, a Web-based annotation platform developed specifically for our project, which is also freely distributed and which integrates a syntactic parser, allowing a semi-automatic annotation. This paper gives a description of the annotation guidelines and the annotation procedure using Contemplata, as well as a statistical description of our corpus, compared with the French Treebank, the largest constituency-based resource for written French.

1 Introduction : projet Temporal@ODIL

La représentation, l'étude et le traitement de la temporalité sont des questions importantes tant pour le Traitement Automatique des Langues Naturelles que pour les sciences du langage. Son étude en corpus nécessite la construction de ressources annotées en temporalité, besoin qui a bénéficié au cours des dix dernières années d'un effort de normalisation dans le cadre du sous-comité TC37/SC4 de l'ISO. Ces travaux ont conduit à la définition d'un standard d'annotation, ISO-TimeML (ISO 2012), qui a été utilisé sur une grande diversité de langues (anglais, italien, coréen, roumain, mandarin, français, etc.) avec des adaptations propres à chaque langue qui sont restées limitées.

Le projet Temporal@ODIL vise précisément à élaborer un nouveau corpus de français annoté en temporalité et se distingue du *French Time Bank* (Bittar et al. 2011), ressource déjà existante, par la nature orale de ses données. Le besoin d'un schéma d'annotation permettant une représentation plus précise de l'empan linguistique des éventualités, c'est-à-dire toutes les mentions dignes d'intérêt : actions, processus, états, etc. (Bach 1981) - nous a conduit à aménager la norme ISO-TimeML tout en cherchant à conserver une cohérence maximale et une compatibilité ascendante avec ce standard (Lefeuvre-Halftermeyer et al., 2016). L'originalité la plus saillante de cette proposition a été de caractériser les unités temporelles (mentions) non pas par leur tête lexicale, comme le requiert ISO-TimeML, mais par le constituant syntaxique qui englobe tous les éléments essentiels à la définition de la mention. L'exemple ci-dessous illustre cette différence de segmentation :

(1) *Paul* [*mange*]$_{u1}$ *une pomme et Pierre* [*mange*]$_{u2}$ *une poire* `ISO-TimeML`
[*Paul mange une pomme*]$_{u1}$ *et* [*Pierre mange une poire*]$_{u2}$ `Temporal@ODIL`

ISO-TimeML définit deux mentions temporelles correspondant à des actions, mais ne les distingue pas syntaxiquement : c'est à chaque fois le même verbe qui caractérise deux mentions différentes. Cette annotation ambiguë est évitée par Temporal@ODIL, puisque l'annotation englobe ici à la fois le sujet et les arguments du verbe. D'un point de vue théorique, la résolution des anaphores temporelles abstraites peut parfois nécessiter la considération de propositions ou tours de parole complets (Zinsmeister et Dipper, 2010).

Si notre schéma d'annotation gagne en précision, ISO-TimeML garde l'avantage de la simplicité de segmentation. On peut toutefois s'interroger sur la facilité que représente, pour un annotateur, une tâche de segmentation à large empan telle que proposée par Temporal@ODIL. L'annotation sera-t-elle fiable ? Peut-on atteindre un bon accord inter-annotateur quant au choix des éléments (arguments sous-catégorisés du verbe par exemple) à intégrer dans chaque mention temporelle ? Pour augmenter la fiabilité de cette étape de segmentation temporelle, nous avons adopté une démarche déjà suivie avec succès par le *Prague Dependency Treebank* (Bejček et Straňák, 2010) : les mentions temporelles sont directement attribuées à des nœuds de la structure syntaxique de l'énoncé. La tâche de segmentation se limite ainsi à la sélection d'un nœud particulier de l'arbre syntaxique considéré. La connaissance de la structure syntaxique de l'énoncé qu'implique ce choix d'annotation peut également faciliter l'annotation des relations temporelles entre mentions. Ainsi, la plupart des relations de subordination temporelle (relation SLINK suivant la norme ISO-TimeML) peuvent être automatiquement inférées à partir des subordinations syntaxiques au sein de l'énoncé.

L'annotation temporelle conduite dans le projet Temporal@ODIL repose ainsi sur la constitution au préalable d'un corpus arboré en constituants. Cet article décrit précisément ODIL_Syntax, le corpus arboré de français oral spontané que nous avons développé. Nous décrivons dans un premier temps les choix d'annotation que nous avons définis, puis la procédure d'annotation réalisée avec Contemplata, un outil permettant de réaliser une annotation syntaxique semi-automatique développée spécifiquement pour le projet. Nous décrivons enfin la ressource réalisée et présentons une analyse distributionnelle en parties du discours que nous comparons avec un corpus de référence du français écrit, le *French Treebank* ou *FTB* (Abeillé et al., 2003; Abeillé et Barrier, 2004).

2 ODIL_Syntax : conventions d'annotation

Afin de permettre une capitalisation des efforts, le corpus ODIL_Syntax suit d'une manière générale les conventions d'annotations définies pour le *FTB*[i]. Ce corpus arboré ne concernant toutefois que le français écrit, nous avons dû procéder à certains ajouts pour représenter en particulier les disfluences orales. La plupart d'entre eux sont compatibles avec les choix d'annotations faits par Abeillé et Crabbé (2013) mais aussi avec ceux du projet Rhapsodie (Lacheret et al. 2014), avec toutefois un degré de granularité moindre. Cette section décrit les principaux ajouts apportés aux conventions du *FTB*.

2.1 Inachèvements

Les inachèvements, tels que définis en linguistique de l'oral (Blanche-Benveniste 1990), correspondent aux situations où le locuteur interrompt subitement son élocution, ou bien commence de manière impromptue un nouvel énoncé qui ne partage aucune cohérence syntaxique avec le tour de parole déjà amorcé. Ils sont annotés explicitement dans le corpus pour permettre leur analyse linguistique, mais également pour les ignorer lors de l'utilisation du corpus à des fins d'apprentissage automatique : ces structures bruitées pourraient en effet conduire à l'élaboration de modèles de langue imparfaits.

Les inachèvements se traduisent par la production d'énoncés intégrant des constituants incomplets. Nous avons choisi de les délimiter au niveau du nœud syntaxique incomplet situé le plus haut possible dans l'arbre des constituants. Ce nœud est alors étiqueté avec la catégorie syntaxique correspondant à l'élément incomplet (ou du moins la plus probable), précédée par une marque spécifique (`$`) rendant compte de cette incomplétude.

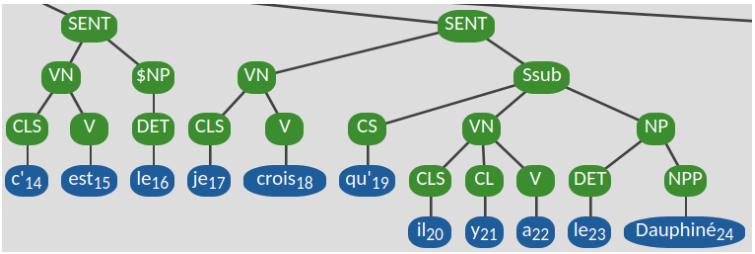

Fig. 1. Exemple d'annotation d'inachèvement : *c'est le- je crois qu'il y a le Dauphiné* [`2AP0292` (OTG)]

Dans l'exemple de la figure 1, l'étiquette `$NP` traduit le fait que le déterminant *le* commence visiblement un groupe nominal inachevé. La marque d'inachèvement est placée le plus haut possible. En effet, le nœud `SENT` ne peut être considéré comme inachevé. Il comprend en effet, de manière régulière, un noyau verbal `VN` dominé par la copule *est* et un groupe nominal inachevé.

L'inachèvement peut intervenir en cours même de production d'un mot. Dans ce cas, si, sa partie du discours (*POS* pour *Part Of Speech*) peut être inférée, on choisit de représenter l'inachèvement au niveau du constituant le plus bas qui domine le POS inachevé, et non pas sur le POS même[ii]. La figure 2 résume cette situation. On y observe en effet une amorce de mot commençant par *g*, mais l'inachèvement est marqué au niveau du groupe nominal `$NP`.

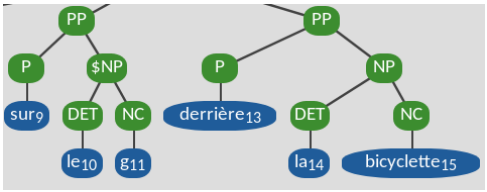

Fig. 2. Exemple d'annotation d'inachèvement : *sur le g- derrière la bicyclette* [`217_C` (ESLO)]

2.2 Entassement paradigmatique : répétitions et réparations

Les répétitions (Ex. 2a), reprises ou autocorrections (Ex. 2b) sont des types de disfluences très fréquentes à l'oral spontané. Elles se traduisent par un entassement paradigmatique, où la progression du discours sur l'axe syntagmatique s'arrête, pour mettre en place un travail de recherche plus ou moins contrôlée de bonne dénomination suivant un paradigme donné. On assiste alors à l'entassement suivant l'axe paradigmatique de plusieurs éléments qui remplissent le même rôle dans l'énoncé (Blanche-Benveniste 1990).

(2a) *C'est c'est madame X* répétition
(2b) *Dans deux minutes trois minutes plutôt* réparation / reformulation
[exemples tirés de `022_00000017` (UBS)]

Le premier élément de l'entassement est appelé reparandum (*deux minutes* dans l'exemple 2b), tandis que le second élément (*trois minutes*) est la réparation de la disfluence. Si certains auteurs (Levelt 1983) ont proposé de représenter ces phénomènes d'une manière analogue aux coordinations, cette approche ne permet pas de rendre compte de toute la richesse de ce dispositif de recherche de dénomination en direct (Blanche-Benveniste 1987). C'est précisément pour rendre compte de la spécificité de ce phénomène que nous proposons, à la suite du projet Rhapsodie, de les annoter avec une étiquette spécifique : `PARA` (pour *entassement PARAdigmatique*). La figure (3) illustre l'annotation de la répétition représentée par l'exemple (2a): on distingue un reparandum *c'est*, placé sous un nœud `PARA`, qui est ensuite répété à l'identique pour former un noyau verbal (`VN`) qui domine précisément le `PARA`.

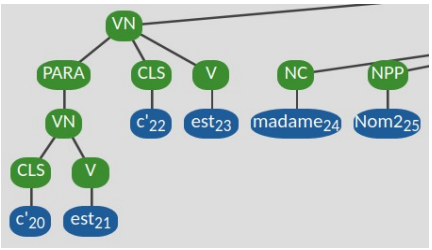

Fig. 3. Exemple d'annotation d'une répétition : *c'est c'est madame Nom2* [`022_00000017` (UBS)]

Cette convention d'annotation est proche de celle adoptée par *Universal Dependencies*[iii], qui rattache – mais dans le cadre d'une analyse en dépendances – le

reparandum à sa reprise. Si le projet *Rhapsodie* a également fait un choix similaire de représentation, la notion d'entassement y est toutefois plus large que dans notre schéma, puisqu'elle est également mobilisée pour par exemple représenter les coordinations (Kahane et al. 2019). De fait, notre schéma d'annotation ne répond qu'à deux des sept types d'entassement envisagés par *Rhapsodie*: `para_disfl` et `para_reform`, les autres types d'entassement n'étant pas spécifiques à l'oral. Ces deux types d'entassement sont regroupés par Kahane et al. (2019) sous la notion d'entassement *de dicto* (i.e. recherche, liée à la production en direct du discours, de la bonne dénomination pour un référent unique), et s'opposent aux entassements *de re* (i.e. les coordinations, qui portent sur plusieurs référents distincts mais qui jouent le même rôle syntaxique dans l'énoncé).

2.3 Clivages, ellipses et incises

Les structures clivées (3), les ellipses (4) et les incises (5), si elles sont très fréquentes en parole spontanée, ne sont pas spécifiques à l'oral. A ce titre, les conventions d'annotation du *FTB* nous permettent de les représenter sans adaptation.

(3) *C'est la rouge que j'ai achetée* clivage
(4) *Moi aussi* ellipse
(5) *Je l'ai vue je pense oui jeudi* incise / parenthèse

Les structures clivées ne feront ainsi l'objet d'aucune représentation spécifique : on annote la structure syntaxique de l'énoncé clivé comme pour tout énoncé ordinaire. De même, les ellipses ne donnent pas lieu à des recommandations particulières : l'annotateur cherchera à construire un arbre de constituants le plus cohérent possible, même si certains éléments sous-catégorisés peuvent alors manquer par rapport aux attentes.

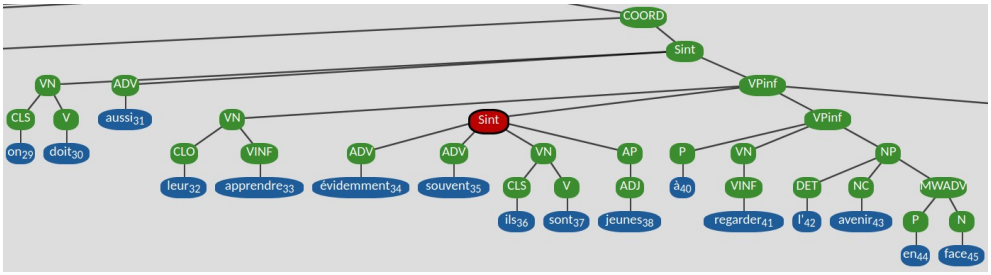

Fig. 4. Exemple d'annotation d'incise : *évidemment souvent ils sont jeunes* [`2AP0292` (OTG)]

Enfin, nous annotons les incises en utilisant le concept d'énoncé interne défini par le *FTB* : l'incise sera représentée par un sous-arbre de constituants ayant pour racine un nœud de type `Sint` (figure 4). L'incise ne partageant aucun lien syntaxique avec l'énoncé principal, elle sera ancrée dans l'arbre de la principale à l'endroit précis où elle intervient dans l'ordre syntagmatique.

2.4 Jeu d'étiquettes utilisé

En dernier lieu, il semble important de préciser le jeu d'étiquettes (POS mais également catégories de syntagmes) utilisé dans nos arbres de constituants. D'une manière générale, nous reprenons ici le jeu d'étiquettes employé pour le modèle du français du *Stanford Parser* (Green et al., 2011). C'est en effet ce parseur qui a été utilisé pour la procédure semi-automatique d'annotation (voir section 3).

Le tableau 1 donne la liste des étiquettes utilisées. On note que l'étiquette `C` est utilisée dans le cas des parties de conjonctions (*parce* et *que*, regroupées par la suite sous le `MWC` *parce que*) et des conjonctions en début de phrase qui n'ont qu'une valeur phatique (par exemple dans *puis alors pas pareil*). On observe également l'absence d'exemples pour `ADJWH`, `ET` et `PREF`, étiquettes connues du *Stanford Parser*, puisqu'aucune occurrence de ces catégories ne figure dans notre corpus.

Tab.1. Jeu d'étiquettes morphosyntaxiques du corpus ODIL

Catégorie	Étiquette	Exemple(s)	Catégorie	Étiquette	Exemple(s)
Adjectif	ADJ	*toute, effarant*	Nom commun	NC	*monsieur, gens*
Adjectif interrogatif	ADJWH		Nom propre	NPP	*Blanc, Loire*
Adverbe	ADV	*pas, enfin*	Ponctuation	PUNC	*?*
Adverbe interrogatif	ADVWH	*quand, combien*	Préposition	P	*au, à*
Conjonction	C	*parce, que, puis*	Préfixe	PREF	
Conj. de coordination	CC	*ou, c'est-à-dire*	Pronom	PRO	*quelqu'un, elle*
Conj. de subordination	CS	*si, que*	Pronom relatif	PROREL	*qui, lesquelles*
Clitique	CL	*y (il y a)*	Pronom interrogatif	PROWH	*quoi, lesquels*
Clitique objet	CLO	*y, me*	Verbe	V	*reviendrez, fait*
Clitique réflexif	CLR	*nous, se*	Impératif	VIMP	*écoutez, voyons*
Clitique sujet	CLS	*on, je*	Infinitif	VINF	*mener, mettre*
Déterminant	DET	*quelques, les*	Participe passé	VPP	*bouleversée, fini*
Déterminant interrogatif	DETWH	*quel*	Participe présent	VPR	*pédalant, appartenant*
Mot étranger	ET		Subjonctif	VS	*aies, suive*
Interjection	I	*euh, oh*			

Ce jeu d'étiquettes est quasiment identique à celui utilisé par Crabbé et Candito (2008) pour la version qu'ils qualifient d'intermédiaire du corpus *FTB,* et qui avait également servi à entraîner un analyseur syntaxique. Les auteurs avaient enrichi le jeu d'étiquettes standard du FTB à l'aide de traits morphosyntaxiques plus discriminants. Quelques différences peuvent être toutefois soulignées avec notre propre jeu d'étiquettes. D'une part, les prépositions et articles contractés sont annotés simplement avec la catégorie `P` dans ODIL_Syntax, et non pas avec le POS `P+D`. D'autre part, nous utilisons un POS générique `CL` pour les clitiques qui ne sont ni sujet (`CLS`), ni objet (`CLO`) ni réflexif (`CLR`). Cette catégorie générique, reprise du *Stanford Parser*, est par exemple utilisée pour certains clitiques explétifs (i.e. non référentiels), tel le *y* de l'expression *il y a* (cf. figure 1).

3 ODIL_Syntax : procédure d'annotation

Afin de simplifier la tâche d'annotation et d'atteindre une certaine fiabilité des données, le corpus ODIL_Syntax a été réalisé suivant une procédure semi-automatique : celle-ci consiste à alterner les recours à un analyseur syntaxique et les étapes de révision manuelle. Cette procédure est réalisée directement sur Contemplata, une plateforme d'annotation spécifiquement développée pour le projet. Cet outil étant diffusé librement comme le corpus, il nous paraît important d'en présenter ici les principales fonctionnalités.

3.1 Contemplata : présentation générale

Contemplata est une plateforme d'annotation basée sur une architecture client/serveur qui permet à l'annotateur de travailler simplement à partir d'un navigateur Web. L'annotateur est ainsi déchargé de toute installation de logiciel sur son ordinateur. Cette architecture permet en outre de proposer des fonctionnalités avancées sans avoir à se préoccuper de considérations techniques, comme par exemple le recours à un analyseur syntaxique ou encore l'adjudication entre plusieurs annotations parallèles.

Un des atouts de Contemplata réside dans ses capacités d'adaptation simplifiée pour toute tâche d'annotation reposant sur une représentation syntaxique en constituants. Ainsi, pour les besoins d'une campagne d'annotation particulière, la simple modification d'un fichier de configuration permet de redéfinir:

- le jeu d'étiquettes défini par les conventions d'annotation,
- dans le cas de l'enrichissement d'un corpus arboré, les types d'entités pertinentes identifiées sur les arbres syntaxiques (dans le cadre du projet : éventualités et expressions temporelles),
- la liste des attributs descriptifs liés à chaque entité ainsi identifiée.

Une autre fonctionnalité saillante de Contemplata réside dans sa capacité à permettre une annotation semi-automatique recourant à un analyseur syntaxique. La spécificité de Contemplata est en effet la cohésion très forte entre les étapes d'analyse automatique et de révision manuelle : notre outil est capable de relancer l'appel à l'analyseur tout en tenant compte des révisions opérées. Celles-ci jouent en effet le rôle de contraintes que l'analyseur se doit de respecter. A l'heure actuelle, deux analyseurs sont intégrés dans la plateforme : *Stanford Parser* et *Disco-Dop* (van Cranenbourg et al. 2016). L'outil permet même d'imposer certaines préférences générales d'analyse sous forme par exemple de contraintes spécifiques de tokenisation ou d'étiquetage morpho-syntaxique. Pour une présentation technique plus fouillée, on se référera à Waszczuk et al. (2020).

3.2 Contemplata : interface annotateur

La figure 4 présente l'interface utilisateur de Contemplata lors du travail d'annotation. Dans cet usage, l'espace de travail affiché par le navigateur web consiste en deux fenêtres d'annotation superposées qui montrent les arbres syntaxiques respectifs de deux tours de paroles qui peuvent être au choix identiques ou différents. Ce partage en deux espaces d'annotation est typiquement dédié à l'annotation de relations quelconques entre nœuds des arbres concernés. Dans le cadre du projet ODIL@Temporal, ces relations correspondent à des relations temporelles, mais on pourrait par exemple imaginer d'utiliser la plateforme pour définir des relations de coréférences ou de toute autre nature. Ce partage de l'espace est également utile pour la révision des annotations : il permet en effet une adjudication entre deux annotations concurrentes. La personne en charge de la révision de l'annotation peut ainsi comparer directement les différences d'annotation, et apporter les modifications qu'elle souhaite à l'une ou l'autre des annotations.

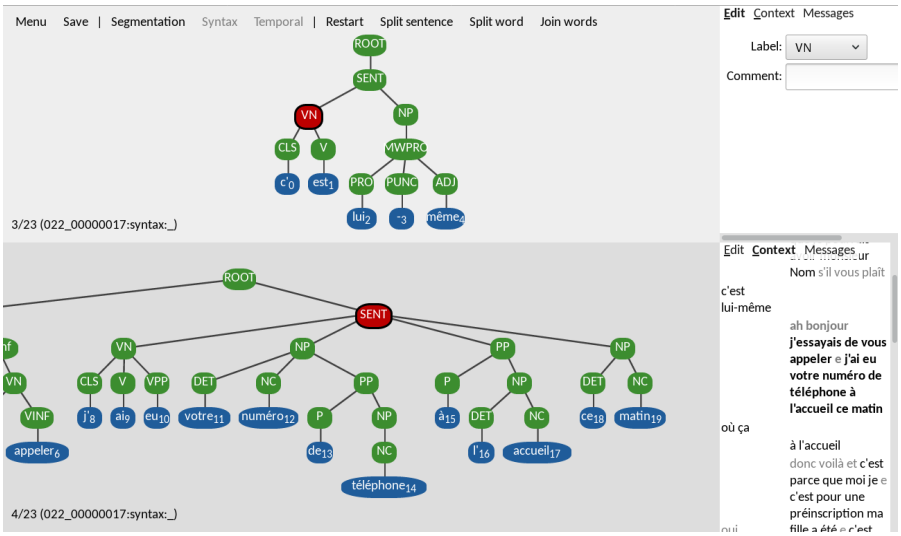

Fig. 5. Interface utilisateur de Contemplata en mode annotation ou adjudication

Comme expliqué plus tôt, Contemplata propose un mode d'annotation semi-automatique combinant un recours à un analyseur automatique et une révision manuelle des sorties de ce dernier. Un grand soin en termes de facilité d'utilisation a ainsi été apporté sur la correction des arbres syntaxiques. L'outil permet ainsi de réaliser de manière purement graphique tout type d'opération de modification d'arbres : ajout, déplacement ou suppression de nœud syntaxique, modification d'étiquette, etc. L'outil n'autorise que des modifications qui respectent la bonne formation des arbres, suivant les contraintes définies dans le guide d'annotation. Un mode « ligne de commande » permet de réaliser des opérations plus complexes comme des opérations d'analyse et d'annotation automatique.

A la droite de chacun des deux espaces d'annotation, on trouve une fenêtre contextuelle qui permet différents types d'opérations :

- le mode `edit` (fenêtre d'annotation supérieure dans la figure 5) permet de renseigner ou modifier les valeurs d'attributs d'annotation du nœud syntaxique sélectionné (en rouge), le noyau verbal (nœud `VN`) dans notre cas. Notons que le fichier de configuration utilisé pour une campagne d'annotation donnée permet de définir des valeurs d'attribut par défaut. Celles-ci sont affectées automatiquement, ce qui allège la tâche de l'annotateur.
- le mode `Context` (fenêtre d'annotation inférieure dans la figure 5) permet d'observer le tour de parole (ou la phrase dans le cas de corpus écrits) auquel correspond l'arbre syntaxique affiché dans son contexte d'occurrence : l'énoncé apparaît en gras dans le fil du discours. Cet affichage contextuel peut aider à désambiguïser certaines annotations et est également utilisé pour se déplacer dans le corpus. On remarque que certains éléments sont grisés : ce sont des éléments (phatiques dans le cas présent) qui n'ont pas été intégrés aux arbres syntaxiques. Là encore, il est possible de réaliser cette opération d'exclusion (et de réinsertion si besoin est) manuellement ou automatiquement. Il est enfin important de noter que cette fenêtre de contexte peut permettre de fusionner plusieurs tours de parole (resp. plusieurs phrases) lorsque le besoin s'en fait sentir[iv] par une simple sélection des tours à concaténer en maintenant la touche `CTRL`.
- mode `Message` – Ce mode sert à afficher les messages d'erreurs envoyés par le serveur.

L'espace respectif accordé à ces quatre zones de travail est totalement modifiable à l'aide des touches ⇨ ⇦ ⇧ ⇩ du clavier. L'annotation d'un seul tour de parole peut ainsi être très confortable d'un point de vue visuel. Notons enfin que pour les campagnes

d'annotation relativement complexes, il est possible de décomposer le travail en différentes sous-tâches, et d'adapter l'interface en fonction de la sous-tâche à réaliser. Le menu en haut de la figure 5 montre précisément que la tâche d'annotation ODIL@Temporal a été décomposée en 3 sous-tâches successives :

- `Segmentation` – regroupement des tours de parole qui auraient été artificiellement découpés lors de la transcription du corpus oral. L'interface présentée en figure 5 correspond justement à la réalisation de cette sous-tâche.
- `Syntax` – réalisation du corpus arboré.
- `Temporal` – caractérisation des entités temporelles et de leur mise en relation.

Cet article se focalise sur la présentation du corpus arboré ODIL_Syntax et laisse de côté la question de l'annotation temporelle qui est encore en cours de réalisation. Nous allons donc désormais décrire la procédure d'annotation syntaxique que nous avons suivie pour réaliser ODIL_Syntax. Pour plus de détails sur l'utilisation pratique de la plateforme `Contemplata`, en particulier pour les rôles de superviseur d'une campagne d'annotation ou bien d'administrateur système, on se référera là encore à Waszczuk et al. (2019).

3.3 ODIL_Syntax : annotation semi-automatique

Comme expliqué précédemment, l'annotation du corpus arboré ODIL_Syntax a été réalisée suivant un processus incrémental combinant des étapes d'analyse automatique et de révision manuelle. Cette stratégie incrémentale allège la charge cognitive de l'annotateur et peut améliorer la fiabilité de l'annotation obtenue. Cinq étapes successives sont réalisées :

- **Nettoyage** (automatique) – exclusion des interjections ou phatiques qui n'intéressent pas notre future annotation temporelle. Des entités telles que les marques de confirmation faibles, les formules d'introduction (*bonjour*) ou d'obligation sociale ont ainsi été exclues des arbres syntaxiques, comme expliqué dans la section 3.2.
- **Première analyse syntaxique** (automatique) – étape réalisée directement à la suite du nettoyage automatique, lors du chargement d'un fichier d'annotation. Elle fait appel, dans le cas présent, au *Stanford Parser*.
- **Révision de la segmentation et des POS** (manuel) – étape où intervient l'annotateur expert pour la révision de la segmentation, rendue nécessaire par le fait que le *Stanford Parser,* entraîné sur un corpus écrit, se base sur le concept de phrase, qui n'est pas opérant sur le français oral. Certains tours de parole sont ainsi parfois composés d'énoncés successifs adjoints, ce qui désoriente complètement l'analyseur automatique. Il peut alors être utile de segmenter le tour de parole en plusieurs pseudo sous-énoncés pour aider l'analyse automatique, sans perdre en précision linguistique. Dans la figure 6, un tour de parole (`ROOT`) est ainsi segmenté manuellement en plusieurs sous-énoncés (`SENT`) indépendants. A l'opposé, cette étape peut également consister à fusionner plusieurs tours de parole qui forment manifestement un seul acte de parole. Lors de cette étape, on demande ensuite à l'annotateur de corriger manuellement les étiquettes morpho-syntaxiques qui auraient été mal identifiées par le *Stanford Parser*.

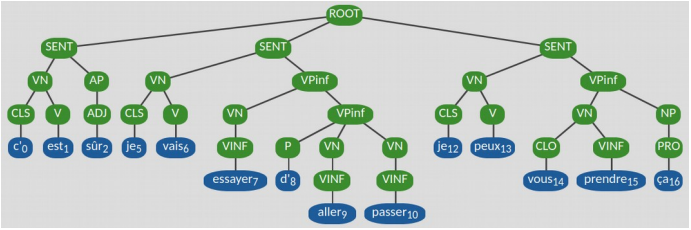

Fig. 6. Segmentation d'un tour de parole en plusieurs énoncés [`2AP0292` (OTG)]

- **Réanalyse** (automatique) – étape qui consiste à appeler à nouveau le *Stanford Parser* en l'obligeant à tenir compte des corrections manuelles déjà effectuées, corrections que l'analyseur prend comme contraintes à respecter (cf. section 3.1). L'expérience montre que cette première révision manuelle est le plus souvent utile à l'analyseur, qui est alors à même de fournir des structures syntaxiques de bonne qualité, bien qu'ayant été initialement développé pour le français écrit.
- **Révision finale** (manuel) – étape lors de laquelle l'annotateur corrige les erreurs subsistantes et vérifie la conformité de l'arbre obtenu avec les conventions d'annotation du *FTB*. C'est lors de cette étape qu'est également gérée l'annotation des disfluences orales telles que les inachèvements et les entassements paradigmatiques. Ces disfluences font l'objet de conventions spécifiques (cf. sections 2.1 et 2.2.) totalement ignorées du *Stanford Parser*. Le recours à l'annotation manuelle est ici nécessaire.

Cette étape d'annotation est enfin suivie d'une étape d'adjudication. Dans le cas présent, le corpus a été annoté par une seule experte, mais révisé ensuite par trois superviseurs afin de garantir une bonne fiabilité de l'annotation. L'estimation de cette qualité ne peut toutefois pas être réalisée par une mesure d'accord inter-annotateur, puisque la procédure d'adjudication a été réalisée par recherche de consensus et non de mise en comparaison.

4 Résultat : description du corpus ODIL_Syntax distribué

Dans cette section, nous caractérisons notre corpus en nous appuyant sur sa constitution, sur la distribution des POS ainsi que sur l'étude statistique des disfluences orales.

4.1 Constitution du corpus

ODIL_Syntax est un corpus visant à représenter différentes variétés (ou registres) de l'oral spontané. Il a ainsi été constitué à partir d'échantillons de corpus créés pour des projets de recherche antérieurs et présentant différents degrés d'interactivité :

- ESLO, qui correspond à des entretiens sociolinguistiques réalisés sur la ville d'Orléans (Baude et Dugua 2011, Eshkol-Taravella et al. 2012) ;
- OTG, qui correspond à des dialogues interactifs en présentiel entre des individus et le personnel d'accueil de l'Office du Tourisme de Grenoble ;
- Accueil_UBS, qui correspond à des dialogues interactifs par téléphone recueillis auprès du standard téléphonique d'une université (Nicolas et al., 2002).

De par leur nature, les entretiens des échantillons ESLO ont un degré d'interactivité plus limité que les dialogues des autres sous-corpus : en effet, les réponses des interviewés s'apparentent tantôt à de courtes répliques, tantôt à des monologues. Une description de ces sous-corpus est donnée dans le tableau 2.

Tab. 2. Description des corpus utilisés dans ODIL_Syntax

Corpus d'origine	Type de dialogue	Nb de mots[v]	Nb de tours de parole	Nb d'énoncés	Nb de dialogues
ESLO	Entretiens	9 663	578	838	3
OTG	Conversations	705	42	71	2
Accueil_UBS	Conversations téléphoniques	1987	166	218	6
TOTAL		12 355	786	1127	11

4.2 Caractérisation morphosyntaxique du corpus

La variété de genres oraux présents dans ODIL nous permet de mesurer l'influence du degré d'interactivité sur la distribution des parties du discours.

Tab. 3. Comparaison de la distribution des POS des 3 sous-corpus d'ODIL

Étiquettes	ESLO	UBS	OTG
Adjectifs	3,99 % (319)	2,08 % (34)	3,32 % (19)
Adverbes	9,99 % (798)	7,42 % (121)	12,06 % (69)
Conjonctions	7,34 % (586)	6,56 % (107)	4,90 % (28)
Clitiques	15,25 % (1 218)	18,09 % (295)	18,71 % (107)
Déterminants	10,95 % (874)	10,36 % (169)	9,09 % (52)
Interjections	0,08 % (6)	0,00 % (0)	0.00 % (0)
Noms	16,43 % (1 312)	20,48 % (334)	15,91 % (91)
Prépositions	11,07 % (884)	9,87 % (161)	10,84 % (62)
Pronoms	4,83 % (386)	4,35 % (71)	4,72 % (27)
Verbes	20,06 % (1 602)	20,78 % (339)	20,45 % (117)

En observant le tableau 3, nous notons qu'il n'y a pas de différence significative entre les entretiens d'ESLO et les dialogues davantage interactifs d'UBS et d'OTG. L'ensemble des POS se répartit soit de manière similaire sur les 3 sous-corpus (notamment les pronoms et les verbes) soit de manière graduelle entre les 3 sous-corpus. Si certaines variations paraissent importantes, elles ne sont cependant pas déterminantes une fois prises globalement : par exemple, la proportion importante des noms dans UBS ne change pas le fait que la catégorie des noms reste une des 3 catégories les plus représentées.

Le degré d'interactivité ne semble donc pas jouer de manière significative sur la distribution des POS. On peut considérer que ces sous-corpus sont comparables et qu'il y a une certaine homogénéité interne au sein du corpus ODIL.

Comme nous l'avons mentionné en section 2.4, ODIL_Syntax a été annoté avec un jeu d'étiquettes similaire à celui du FTB, nous permettant ainsi de le mettre en regard de ce corpus arboré de l'écrit pour en étudier les différences. Il est entendu que la modalité seule ne peut pas tenir compte des différences que nous pourrons observer et que les registres ont aussi un impact non négligeable. A ce propos, Dewaele (2001:184) écrit ceci avec justesse : « Biber (1988) conclut qu'il n'existe pas de dimension unique permettant d'expliquer toute la variance entre les registres. Il constate aussi qu'aucune dimension ne permet de saisir toutes les différences entre l'oral et l'écrit mais que, malgré tout, les conversations spontanées apparaissent comme l'oral stéréotypique alors que la prose académique représente l'écrit stéréotypique. »

Par ailleurs, sans être un corpus représentatif de l'oral dans toute sa diversité, ODIL offre une collection de données orales présentant des degrés de spontanéité et d'interactivité variés et portant sur des thématiques plus ou moins larges. Il s'agit d'une des rares ressources de français oral annoté en syntaxe. Il nous semble donc intéressant de la caractériser par rapport à d'autres ressources orales ou écrites disponibles. Outre les distributions calculées sur le FTB, nous empruntons à Dewaele (2001) la distribution en POS des 2 corpus de français oral (OR_informel_F et OR_formel_F) et des 2 corpus de français écrit (ECR_discours_F et ECR_journal_F) qu'il a étudiés. La comparaison de notre corpus avec ces différentes contributions nous permettra de mesurer l'apport de la ressource

que nous proposons, et sa contribution à enrichir l'offre de ressources annotées sur une modalité, l'oral spontané, encore peu disponible sous forme de corpus arboré.

La distribution des POS sur l'ensemble du corpus ODIL a été mise en perspective avec celle du FTB, mais aussi avec celles de Dewaele (2001) dans le tableau 4. Pour ce faire, nous avons dû rassembler les clitiques (`CL`, `CLO`, `CLR`, `CLS`) et les pronoms (`PRO`, `PROREL`, `PROWH`) puisque Dewaele ne distingue pas ces deux catégories. Nous notons également que la catégorie *adjectifs* de Dewaele comprend les "adjectifs" démonstratifs et possessifs de la grammaire traditionnelle alors que nous rangeons ces derniers parmi les déterminants. L'étiquette `DET` que nous utilisons ne nous permet donc pas de faire une distinction automatique des adjectifs démonstratifs et possessifs. En revanche, ils sont bien sous-catégorisés pour le FTB, mais ce transfert ne se traduit que par un glissement de l'ordre de 2% des distributions (9,17% d'adjectifs et 18,07% de déterminants) qui n'est pas de nature à modifier les conclusions de notre étude.

Tab. 4. Comparaison de la distribution des POS dans ODIL et FTB, ainsi que dans Dewaele (2001).

Étiquettes	ODIL	OR informel F	OR formel F	ECR discours F	ECR journal F	FTB
Adjectifs	3,65 % (372)	7,6 %	9,5 %	7,9 %	16,4 %	7,31 % (36 243)
Adverbes	9,70 % (988)	11,5 %	10,9 %	9,3 %	4,2 %	4,65 % (23 045)
Conjonctions	7,08 % (721)	5,1 %	7,8 %	7,2 %	4,3 %	3,70 % (18 362)
Déterminants	10,72 % (1 093)	10,6 %	11,1 %	14,7 %	17,6 %	19,93 % (98 779)
Interjections	0,06 % (6)	2,5 %	1,8 %	0.00 %	0,00 %	0.01 % (70)
Noms	17,05 % (1 737)	16,2 %	17,1 %	20,2 %	27,7 %	27,33 % (135 451)
Prépositions	10,86 % (1 107)	7,8 %	9,9 %	11,6 %	15,0 %	17,33 % (85 917)
Pronoms (clitiques inclus)	20,68 % (2 107)	18,6 %	12,7 %	11 %	3,6 %	5,91 % (29 302)
Verbes	20,20 % (2 058)	20 %	18,9 %	17,7 %	10,7 %	13,82 % (68 523)

Le corpus ECR_journal_F est, de par sa nature, semblable au corpus FTB puisqu'il s'agit aussi d'un écrit journalistique. Cette similarité se retrouve dans la distribution des POS de façon remarquable, mis à part pour les adjectifs qui sont beaucoup plus nombreux dans ECR_journal_F, même après ajout des "adjectifs" démonstratifs et possessifs. De l'autre côté, on remarque que la distribution des POS dans ODIL le place clairement avec les corpus oraux de Dewaele mais ne le rapproche d'aucun d'entre eux de manière claire.

En regardant de plus près la distribution des différentes catégories dans ODIL et le FTB, nous remarquons des différences importantes dont nous justifions la significativité par des tests[vi]. D'après Dewaele (2001), le choix de la modalité orale déchargerait le locuteur de l'usage accru des substantifs et des prépositions fait par le scripteur[vii] et le pousserait à

utiliser davantage de verbes (conjugués), d'adverbes et de pronoms, catégories plus à même à ancrer son discours dans le contexte spatio-temporel. C'est aussi ce que nous observons.

Les **noms communs** sont ainsi 10 % plus nombreux à l'écrit qu'à l'oral. On montre que cette différence d'usage entre les corpus ODIL et FTB est statistiquement significative selon un test bilatéral de Wilcoxon-Mann-Whitney[viii] (test bilatéral de U = 0 = Ucritique(0,05)). On peut en effet rejeter l'hypothèse d'égalité des deux distributions avec une confiance de α niveau de confiance de α = 0,05 (5%), différence également confirmée à un niveau de confiance de α = 0,01 (1%) si on ajoute les études de Müller (1985) et Dewaele (2001) pour lesquels U = 0 < Ucritique(0,01) = 0. De même, les **prépositions** sont 7 % plus nombreuses dans le FTB que dans ODIL. On montre que cette différence de fréquence d'occurrences est statistiquement significative à un niveau de confiance de α = 0,05 (5%). Par la même méthodologie que celle décrite pour les noms, on obtient une valeur de test bilatéral de U = 0 = Ucritique(0,05)). Müller (1985) et Dewaele (2001) n'ayant pas étudié spécifiquement les prépositions, nous n'avons pas d'éléments de comparaison avec leurs études.

De l'autre côté, tout comme Moscovici et Humbert (1960), nous observons que le **ratio verbes/nom communs** est plus élevé à l'oral qu'à l'écrit, observation justifiée selon Halliday (1989) et Biber et al. (1998:69) par le caractère plus dynamique de l'oral. Cette différence de fréquence d'occurrences des verbes entre les corpus ODIL et FTB est statistiquement significative (confiance de α = 0,05 (5%) et test bilatéral de U = 0 = Ucritique(0,05)). On retrouve cette différence significative avec le registre général du français écrit journalistique : si l'on ajoute au corpus FTB les études de Müller (1985) et celle de Dewaele (2001) sur le corpus ECR_journal_F, on observe également un rejet de l'hypothèse d'identité par un test de Wilcoxon-Mann-Whitney U = 0 = Ucritique(0,05). Pour ce qui est des **adverbes,** nos chiffres nous rapprochent une nouvelle fois de Moscovici et Humbert (1960) puisque nous retrouvons également que les adverbes deux fois plus nombreux à l'oral qu'à l'écrit. Cette différence est également statistiquement significative (niveau de confiance de α = 0,05 (5%), test bilatéral de U = 0 = Ucritique(0,05)). Cette différence est confirmée en ajoutant les études de Müller (1985) et (Dewaele 2001) : U = 1 < Ucritique(0,05) = 2. Enfin, la prédominance des **pronoms** à l'oral est telle qu'aucun test n'est nécessaire pour démontrer que la différence avec l'écrit est significative. L'explication se trouve de manière évidente du côté des pronoms personnels de première et de deuxième personne qui comptent pour près du tiers de l'ensemble des pronoms dans ODIL. Redecker (1984) justifie cette prédominance par le fait que l'oral se situe du côté du pôle implication (*involvement* chez Chafe (1982) et Tannen (1982)) avec donc plus d'autoréférences à l'oral (notamment du pronom personnel *je*) tandis que Mazzie (1987) propose que l'oral est moins explicite que l'écrit.

Pour mieux comprendre maintenant les caractéristiques d'ODIL en tant que corpus oral, nous nous intéressons maintenant à l'analyse des disfluences spécifiques à l'oral.

4.3 Structures de l'oral

Dans la description du corpus H-H oral en anglais (TRAINS corpus), Heeman et Allen (1999) ont observé la présence de 2399 réparations/répétitions pour 6163 tours de parole. Notre corpus est davantage affecté par les disfluences orales puisque nous comptons 366 entassements paradigmatiques et 186 inachèvements : à peine moins de la moitié des tours

de parole et moins d'un tiers des énoncés contiendrait donc un entassement, tandis que l'on trouverait un inachèvement dans 1 tour de parole sur 5 et 1 énoncé sur 6.

Tab. 5. Répartition des entassements dans ODIL, par catégorie

Catégories	Taux de PARA	Catégories	Taux de PARA
Adjectif	1 (0,27%)	Syntagme adjectival (AdP)	2 (0,54%)
Adverbe	12 (3,24%)	Syntagme adverbial (AP)	3 (0,81%)
Conjonction	16 (4,32%)	Syntagme coordonné (COORD)	2 (0,54%)
Clitique	40 (**10,81%**)	Syntagme nominal (NP)	52 (**14,05%**)
Déterminant	45 (**12,16%**)	Syntagme prépositionnel (PP)	20 (5,41%)
Nom	6 (1,62%)	Proposition interne (Sint)	4 (1,08%)
Préposition	44 (**11,89%**)	Proposition relative (Srel)	8 (2,19%)
Pronom	4 (1,08%)	Proposition subordonnée (Ssub)	11 (2,97%)
Verbe	6 (1,62%)	Noyau verbal (VN)	68 (**18,38%**)
		Énoncés (SENT)	20 (5,41%)

Le tableau 5 nous montre que près d'un tiers des entassements sont concentrés sur deux syntagmes, le noyau verbal (voir figure 3, *c'est c'est madame Nom2*) et le syntagme nominal (Ex. 6), qui sont des noyaux de sens importants pour la compréhension des énoncés. Il n'y a donc rien d'étonnant au fait de trouver davantage de reprises et de recherches de dénomination à ces niveaux-là.

(6) (PARA NP *un chemin de fer*) NP *un tout petit chemin de fer* [217_C (ESLO)]
(7) NP (PARA DET *la*) DET *notre* NC *collègue* [1AP0507 (OTG)]

Souvent également, l'entassement ne s'initie pas une fois le constituant complètement réalisé, mais simplement amorcé. Dans ce cas, l'entassement PARA sera donc marqué au niveau d'un POS, par exemple le déterminant en (7). On observe donc qu'un second tiers des entassements porte sur les prépositions, les déterminants ainsi que les clitiques (dont 80 % sont sujets). Il s'agit donc cette fois de reprises en début de syntagme, ce qui confirme que le constituant a une réalité linguistico-cognitive (Blanche-Benveniste 1990).

Tab. 6. Répartition des inachèvements par type de syntagme

	ESLO	UBS	OTG	TOTAL ODIL
$AdP	0 (0%)	0 (0%)	0 (0%)	0 (0%)
$AP	2 (0,94%)	0 (0%)	0 (0%)	2 (0,79%)
$COORD	12 (5,66%)	0 (0%)	0 (0%)	12 (4,76%)
$NP	44 (**20,75%**)	6 (**19,35%**)	4 (**44,44%**)	54 (**21,43%**)
$PP	18 (8,49%)	1 (3,23%)	1 (**11,11%**)	20 (7,94%)
$Sint	11 (5,19%)	0 (0%)	0 (0%)	11 (4,37%)
$Srel	13 (6,13%)	2 (**6,45%**)	0 (0%)	15 (5,95%)
$Ssub	30 (**14,15%**)	4 (**12,90%**)	0 (0%)	34 (**13,49%**)
$VN	31 (**14,62%**)	6 (**19,35%**)	0 (0%)	37 (**14,68%**)
$VPinf	0 (0%)	2 (6,45%)	0 (0%)	2 (0,79%)
$VPpart	1 (0,47%)	0 (0%)	0 (0%)	1 (0,40%)
$SENT	50 (**23,58%**)	10 (**32,26%**)	4 (**44,44%**)	64 (**25,40%**)
TOTAL	**212**	**31**	**9**	**252**

D'après le tableau 6, les inachèvements se répartissent quant à eux autour des énoncés (Ex. 8) pour un quart d'entre eux, puis des syntagmes nominaux (Ex. 9a et 9b), des noyaux verbaux (Ex. 10) ainsi que des propositions subordonnées (Ex. 11).

(8) `$SENT` *nous on a été juste / on a été en Alsace* [`217_C` (ESLO)]
(9a) `$NP` *six / à six heures* [`215_C` (ESLO)]
(9b) `$NP` *l'é / l'évasion* [`007_C-1` (ESLO)]
(10) `$VN` *je / c'est pour une préinscription* [`22_00000017` (UBS)]
(11) `$Ssub` *parce que / comment vous expliquer* [`007_C-1` (ESLO)]

Les syntagmes nominaux inachevés qui sont repris par le locuteur peuvent être suivis d'une réparation (Ex. 9a), mais aussi correspondre à la répétition d'un nom incomplet (Ex. 9b). En effet, comme nous l'avons vu en section 2.2, l'incomplétude des mots n'est pas marquée au niveau du POS mais au niveau du constituant le plus bas le dominant.

Au niveau des autres syntagmes et des énoncés, il s'agit le plus souvent d'un faux départ : dans l'exemple (10), alors que nous avons un début de phrase personnelle avec le clitique « je », le locuteur part finalement sur une phrase impersonnelle avec un présentatif. Une bonne partie des énoncés inachevés se trouvent en effet être des présentatifs dont il manque le complément (*c'est...*, *il y a...*). Nous notons enfin que si certains de ces énoncés sont effectivement inachevés et suivis d'un énoncé complètement différent, d'autres font en réalité partie du procédé d'entassement et sont donc des répétitions ou des réparations. Dans ce cas, l'énoncé inachevé est le reparandum, comme le montre la figure 7.

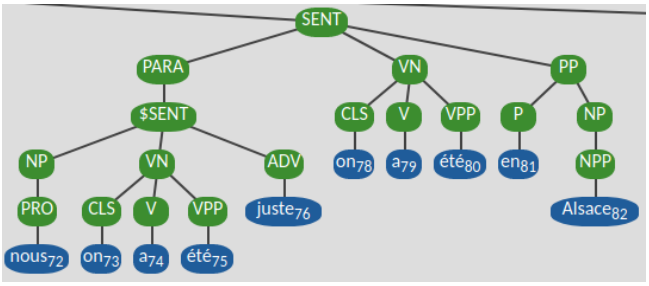

Fig. 7. Inachèvement dans un entassement paradigmatique (illustration de l'Ex.8) [`217_C` (ESLO)]

5 Conclusion

Nous avons présenté un corpus d'oral spontané annoté en arbres de constituants. L'étude de l'annotation syntaxique a non seulement permis de caractériser syntaxiquement notre corpus, mais aussi de montrer qu'ODIL était statistiquement et significativement différent d'un corpus d'écrit journalistique. Si ce registre particulier a été choisi pour comparaison, c'est parce qu'il s'agit du seul treebank annoté en constituants existant pour l'heure pour le français. Les disfluences de l'oral étant annotées, notre corpus permet également une analyse syntaxique approfondie des entassements et des inachèvements. Par ailleurs, l'annotation a été réalisée avec un outil libre adaptable à tout autre corpus de l'oral et à toute autre tâche d'annotation sémantique basée sur des arbres de constituants. Enfin, ODIL_Syntax est un corpus libre distribué sur la plateforme ORTOLANG[ix] sous licence Creative Commons : CC-BY-SA pour les échantillons OTG et Accueil_UBS, et CC-BY-SA-NC pour les extraits d'ESLO.

Références

Abeillé, A. et Barrier, N. (2004). Enriching a French treebank. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Abeillé, A., Clément, L., et Toussenel, F. (2003). Building a treebank for french. Treebanks. *Text, Speech and Language Technology*, 20, pp. 165–187.

Abeillé, A. et Crabbé, B. (2013). Vers un treebank du français parlé. In *20ème conférence du Traitement Automatique du Langage Naturel (TALN'13)*, Sables d'Olonne, France.

Bach, E. W. (1981). Time, Tense, and Aspect: An Essay in English Metaphysics. In *Radical Pragmatics*, pp. 63–81.

Baude, O. et Dugua, C. (2011). (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ?, *Corpus*, 10, pp. 99–118.

Bejček, E. et Straňák, P. (2010). Annotation of Multiword Expressions in the Prague Dependency Treebank. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, 44(1-2), pp. 7–21.

Biber, D., Conrad, S. et Reppen R. (1998). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bittar, A., Amsili, P., Denis, P., et Danlos, L. (2011). French TimeBank: an ISO-TimeML Annotated Reference Corpus. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers* (ACL'2011), pp. 130–134, Portland, USA.

Blanche-Benveniste, C. (1987). Syntaxe, choix de lexique, et lieux de bafouillage. *DRLAV. Documentation et Recherche en Linguistique Allemande Vincennes*, 36(1), pp. 123–157.

Blanche-Benveniste, C. (1990). *Le français parlé : études grammaticales*. CNRS Edition, pp. 17-38.

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. In: D. Tannen (ed.), *Spoken and Written Language: Exploring Orality and Literacy*. Norwood : Ablex, pp. 35–53.

Crabbé, B. et Candito, M. (2008). Expériences d'analyse syntaxique statistique du français. *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles - TALN'08*, pp. 44–54, Avignon, France.

Dewaele, J. M. (2001). Une distinction mesurable: corpus oraux et écrits sur le continuum de la deixis. *Journal of French Language Studies*, 11(2), pp. 179–199.

Eshkol-Taravella, I., Baude, O., Maurel, D., Hriba, L., Dugua, C., et Tellier, I. (2011). Un grand corpus oral "disponible" : le corpus d'Orléans 1 1968-2012. *Traitement Automatique des Langues*, 53(2), pp. 17–46.

Green, S., de Marneffe, M.-C., Bauer, J., et Manning, C. D. (2011). Multiword expression identification with tree substitution grammars: A parsing tour de force with French. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pp. 725–735, Stroudsburg, PA, USA.

Halliday, M. A. K. (1989). *Spoken and Written Language*. Oxford: Oxford University Press.

Heeman, P.A., Allen, J.F. (1999). Speech repairs, intentional phrases and discourse markers : modeling speaker's utterances in spoken dialogue. *Comptutational Linguistics*, 25(4), pp. 527–572.

ISO. (2012). Language Resource Management - Semantic Annotation Framework (SemAF) - Part 1: Time and Events. ISO 24617-1:2012. International Organization for Standardization.

Kahane, S., Pietrandrea, P., et Gerdes, K. (2019). The annotation of pile structures. A. Lacheret-Dujour, et al. (ed), *Rhapsodie: A prosodic and syntactic treebank for spoken French*, pp. 65–95. John Benjamins.

Lacheret, A., Kahane, S., Belião, J., Dister, A., Gerdes, K., Goldman, J.-P., Obin, N., Pietrandrea, P., et Tchobanov, A. (2014). Rhapsodie: a Prosodic-Syntactic Treebank for Spoken French. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, pp. 295–301, Reykjavik, Iceland.

Lefeuvre-Halftermeyer, A., Antoine, J.-Y., Couillault, A., Schang, E., Abouda, L., Savary, A., Maurel, D., Eshkol-Taravella, I., et Battistelli, D. (2016). Covering various Needs in Temporal Annotation: a Proposal of Extension of ISO TimeML that Preserves Upward Compatibility. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'2016)*, Portoroz, Slovenia.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1), pp. 41–104.

Mazzie, C. A. (1987). An experimental investigation of the determinants of implicitness in spoken and written discourse. *Discourse Processes*, 10, pp. 31–42.

Moscovici, S. et Humbert, C. (1960). Etudes sur le comportement verbal. Langage oral et langage écrit. *Psychologie française*, 3, pp. 175–183.

Müller, B. (1985). *Le Français d'aujourd'hui*. Paris: Klincksieck.

Nicolas, P., Letellier-Zarshenas, S., Schadle, I., Antoine, J.-Y., et Caelen, J. (2002). Towards a large corpus of spoken dialogue in French that will be freely available: the "Parole Publique" project and its first realisations. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'2002)*, Las Palmas, Canary Islands.

Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language*, 58, pp. 1–21.

van Cranenburgh, A., Scha, R., et Bod, R. (2016). Data-oriented parsing with discontinuous constituents and function tags. *Journal of Language Modelling*, 4(1), pp. 57–111.

Waszczuk, J., Wang, I., Antoine, J.-Y., et Halftermeyer, A. (2020). Contemplata, a Free Platform for Constituency Treebank Annotation. *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC'2020)*, Marseille, France.

Zinsmeister, H. et Dipper, S. (2010). Towards a Standard for Annotating Abstract Anaphora. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2010)*, pp. 54–59, Valletta, Malta.

[i]Voir le guide des annotations morpho-syntaxiques et celui de l'annotation en constituants disponibles sur http://ftb.linguist.univ-paris-diderot.fr/treebank.php?fichier=documentation dans la section « Choix d'annotation ».

[ii]Il s'agit ici d'un choix purement pragmatique qui peut être contesté. Il se trouve simplement que, dans notre stratégie d'annotation semi-automatique, les analyseurs syntaxiques ne sont pas à même de prendre en compte des marques d'incomplétudes au niveau des POS.

[iii]Le modèle générique *Universal Dependencies* utilise ainsi un lien appelé reparandum : https://universaldependencies.org/fr/dep/reparandum.html

[iv]Nous définissons un tour de parole dès qu'il y a changement de locuteur. Ainsi, si un locuteur et interrompu par son interlocuteur mais poursuit tout de même son élocution, sa prise de parole, correspondant potentiellement à un seul acte de langage, sera scindée en 2 tours de parole. Pour les besoin de l'annotation syntaxique, nous pouvons ainsi fusionner ces deux tours au niveau des arbres, même si la segmentation originale est conservée.

[v]Si le nombre total de mots pour l'ensemble du corpus brut est bien de 12 355 mots, nous notons toutefois que l'exclusion des phatiques en pré-traitement réduit à 10 294 le nombre de mots annotés (8091 pour ESLO, 572 pour OTG et 1631 pour Accueil_UBS).

[vi]Les tests ont été effectués sur un ensemble de 5 sous-corpus pour ODIL (ESLO1, ESLO2, ESLO3, OTG et UBS) et sur un ensemble de 3 sous-corpus pour le corpus FTB. À des fins de reproductibilité des expériences, nous tenons à disposition des personnes intéressées les données source qui nous permettent de calculer nos scores de significativité (contact : Jean-Yves.Antoine@univ-tours.fr).

[vii]C'est aussi le cas des articles et des adjectifs, mais pour les raisons données dans l'article, nous ne pouvons comparer ces catégories à nos déterminants et nos adjectifs.

[viii]Remarquons au passage que (Dewaele 2001) retrouve lui aussi qu'à l'exception des verbes, la proportion de toutes les catégories grammaticales qu'il a étudiées sur des corpus variés varie de façon significative entre l'écrit et l'oral. Cette hypothèse est validée par un test paramétrique t de Student. Compte-tenu de la taille restreinte de nos corpus, nous préferons étudier dans cette étude la significativité à l'aide d'un test non paramétrique (Wilcoxon-Mann-Whitney).

[ix]https://www.ortolang.fr/market/corpora/odil