Le projet ORFÉO : un corpus d'étude pour le français contemporain

The ORFEO project: a study corpus for contemporary French

Christophe Benzitoun, Jeanne-Marie Debaisieux et Henri-José Deulofeu

---

Édition électronique

URL : <http://journals.openedition.org/corpus/2936>

DOI : 10.4000/corpus.2936

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Édition imprimée

Date de publication : 15 octobre 2016

ISSN : 1638-9808

Référence électronique

Christophe Benzitoun, Jeanne-Marie Debaisieux et Henri-José Deulofeu, « Le projet ORFÉO : un corpus d'étude pour le français contemporain », *Corpus* [En ligne], 15 | 2016, mis en ligne le 15 janvier 2017, consulté le 10 décembre 2020. URL : <http://journals.openedition.org/corpus/2936> ; DOI : <https://doi.org/10.4000/corpus.2936>

Ce document a été généré automatiquement le 10 décembre 2020.

© Tous droits réservés

Le projet ORFÉO : un corpus d'étude pour le français contemporain

The ORFEO project: a study corpus for contemporary French

Christophe Benzitoun, Jeanne-Marie Debaisieux et Henri-José Deulofeu

---

1. Introduction

- 1 Le projet ORFÉO (Outils et recherches sur le français écrit et oral) est un projet financé par l'Agence nationale de la recherche (ANR 12-CORP-0005) et qui a été retenu dans le cadre de la campagne Corpus, données et outils de la recherche en sciences humaines et sociales 2011. Le projet a démarré en février 2013. Dans cet article, nous présenterons les objectifs généraux du projet, le traitement en cours des données orales (collecte, harmonisation, métadonnées), l'état du travail sur l'enrichissement des données par des annotations syntaxiques et la plate-forme d'interrogation. Dans le texte ci-dessous, nous abordons uniquement la composante orale du projet ORFÉO.

2. Les objectifs généraux du projet

- 2 ORFÉO a pour objectif de mener différentes études comparatives sur des données de genres variés (notamment écrit vs oral) sur un corpus préalablement outillé. La constitution du corpus, et notamment des données secondaires, est la tâche la plus importante. Les corpus que nous rassemblons sont des corpus existants libres de droits ou mis à disposition par les ayants droit. Les données secondaires sont de nature diverse : métadonnées, transcriptions harmonisées, alignement texte et son, annotations morphosyntaxiques (PoS), syntaxiques, sémantiques, conversationnelles, prosodiques, segmentations en unités « élémentaires » de texte (macro-syntaxe). Corpus et données seront accessibles au travers d'une plate-forme d'interrogation permettant une sélection par les métadonnées et des recherches à l'aide de requêtes simples (chaîne de caractères ou expressions régulières) ou complexes (requêtes sur les différents niveaux d'annotation). Les utilisateurs seront guidés dans la sélection et

l'utilisation des outils par l'accès à des analyses pilotes relevant de différents domaines de la linguistique : morphosyntaxe, sémantique, analyse du discours et des interactions. La plate-forme permettra ainsi un accès unique à un corpus d'étude comportant plusieurs millions de mots (3M à l'oral et 6M à l'écrit).

- 3 L'enrichissement des données proposées par des annotations morphosyntaxiques et en dépendances permettra d'interroger sur des structures complexes qu'un accès par concordancier basique ou par expressions régulières simples ne permet pas d'atteindre de façon satisfaisante. Un travail de thèse (Bérard, 2012) portant sur les structures dites « de contrôles à distance » telles que :
  - (1) **qu'est-ce que** vous voulez que j'y **fasse**
  - (2) l'hypothèse **que** je crois qu'on **peut avancer**
  - (3) c'est **où** qu'il faut qu'elle **se présente**
- 4 a montré que les résultats pertinents d'interrogation d'un corpus de 2M de mots sont multipliés par six grâce au recours à des données annotées. Au-delà des progrès envisagés dans l'analyse, le projet relève d'un enjeu politique majeur pour la linguistique de corpus. Il s'agit, d'une part, de mutualiser les résultats de recherches antérieures, conçus dans le cadre des ANR Rhapsodie, Annodis et Decoda et, d'autre part, de donner à la linguistique de corpus un outil de travail conforme aux standards internationaux.

3. Les partenaires du projet

- 5 Le projet est rendu possible par la collaboration de 7 laboratoires français : LATTICE (U. Paris 3/Ens.-UMR 8094), MoDyCo (Université Paris Ouest Nanterre La Défense - UMR 7114), ATILF (U. de Lorraine-UMR 7118), LORIA (CNRS -UMR 750), LIF (Université Aix-Marseille-UMR 7279), ICAR (U. Lyon 2-UMR5191), CLLE-ERSS (U. Toulouse-UMR 5263) et de chercheurs étrangers (Suisse, Belgique, Japon). La complémentarité des recherches impliquant des linguistes et des informaticiens aboutit à une répartition des tâches fonctionnelle au regard de la chaîne de traitement des données :
  - La mise en forme des ressources s'effectue au sein de l'ATILF (sous la responsabilité de C. Benztoun et d'E. Jacquy, avec la collaboration de L. Bérard), qui est chargé de l'unification des formats (données + métadonnées), de l'uniformisation des conventions de transcription et de leur correction.
  - L'alignement automatique au phonème des données orales relève du Loria (resp. C. Cerisara) et permettra des analyses prosodiques fines. Deux outils ont été développés : le LASTAS (Loria Automatic Speech-Text Alignment Software) dédié au traitement de précision de courts segments et JTrans dédié au traitement de longs fichiers de parole.
  - Les corpus oraux n'étant pas ponctués, un premier logiciel de calcul de frontières d'énoncé a été développé au sein de Modyco (Wang, 2013, Wang *et al.*, 2014), dans le cadre d'un stage de travail de master. Mais, compte tenu de la grande diversité des types de texte (*cf. infra*), l'équipe a dû revenir à un système de segmentation manuelle. Cette segmentation est nécessaire au parseur syntaxique qui ne peut calculer de dépendances qu'au sein de segments clairement délimités.
  - Les annotations morphosyntaxiques (PoS) et les annotations en dépendances, inspirées en partie de l'annotation en dépendance développée dans le cadre du projet Rhapsodie (ANR Rhapsodie, 2008-2012, sous la direction d'A. Lacheret) sont le résultat

d'une étroite collaboration entre les linguistes et les informaticiens de l'équipe TALEP (F. Bechet, A. Nasr, Carlos Ramisch, J. Deulofeu, et A. Valli), au sein du LIF.

- 6 L'ensemble de ces outils devraient être testés dans le cadre d'études pilotes développées au sein des laboratoires CLAPI et CLLE-ERSS, portant sur des phénomènes linguistiques sensibles aux effets de genre : les formes non canoniques de « noyaux macrosyntaxiques », la famille des constructions « modales » et les marqueurs d'attitude, la famille des constructions avec réalisations de place syntaxique par listes (p. ex. : *j'ai vu un oiseau une sorte de pélican*) ou avec « effet deux points » (p. ex. : *il y a un truc chez Marie elle est renversante*) et les structures de *clause linking*. Dans ce cadre, certaines sous-parties du corpus doivent être enrichies d'annotations macrosyntaxiques (F. Sabio), d'annotations en relations de discours (M. Ho-dac), en références et coréférences (F. Landragin) et d'annotations « interactionnelles » (V. Traverso).

4. Le traitement des données orales

- 7 La partie orale de la base ORFÉO provient de la mutualisation d'un ensemble de corpus, déjà diffusés pour certains. La totalité des transcriptions était donc existante avant le lancement du projet, mais devait être harmonisée et complétée. Le résultat constitue une large couverture en termes de situations de parole : réunions de travail, entretiens, interviews, réunions publiques, contes, récits, conversations téléphoniques, etc. Au total, plus de 2 000 locuteurs sont représentés dans les enregistrements, provenant de trois pays francophones : Belgique, France et Suisse. Différentes étapes ont été suivies entre la réception des données et leur future diffusion.
- 8 Le premier problème auquel l'équipe a été confrontée réside dans l'hétérogénéité des formats des données collectées (txt, doc, xml, etc.) qui ont dû être transformés en un format unique : Transcriber (.trs). Certains formats ont été par ailleurs conservés pour l'alignement automatique (voir *infra*). Les conventions de transcription étaient également fort hétérogènes. Les transcriptions ont dû être uniformisées afin de permettre l'interrogation de l'ensemble. La convention retenue est celle de l'orthographe standard, qui facilite la lecture et les annotations automatiques. La plupart des systèmes produisent en effet de meilleurs résultats sur des données linguistiques respectant les conventions orthographiques standard. Un système de balise a permis de séparer texte et commentaires (événements extralinguistiques, rires, commentaires, pauses, etc.).
- 9 Le travail le plus important a été celui de correction des transcriptions qui a porté sur une grande partie de la base. Un grand nombre de transcriptions ont été revues par deux personnes différentes. Ce travail est extrêmement chronophage (plus d'une heure pour vingt-cinq minutes de transcription). Il ne garantit pas l'absence d'erreur mais le haut niveau de qualité des transcriptions proposées. Parallèlement à la phase de correction, une partie manuelle du travail d'homogénéisation a été menée. Une partie des corpus a également fait l'objet d'un alignement manuel texte/son afin de faciliter l'alignement ultérieur. La tâche de correction a été couplée au repérage des segments à anonymiser et à l'enrichissement des métadonnées par la rédaction d'un bref résumé du contenu. L'ensemble des transcriptions a été hébergé sur un système de versionnage et mis à jour au fur et à mesure de l'avancement du travail.

- 10 L'équipe a également développé un outil permettant de soustraire automatiquement des fichiers sonores les passages non transcrits qui peuvent se situer en début, en fin, mais aussi au cours de l'enregistrement. Après avoir repéré les sections non transcrites (marquées préalablement par des sections « nontrans » dans le logiciel Transcriber), l'outil supprime les passages en question du fichier son, insère une balise correspondant à la durée retranchée et recalcule automatiquement l'alignement texte/son pour la suite de la transcription. Au final, les fichiers sons correspondent exactement aux passages transcrits. La démarche évite le risque de diffusion de données sensibles non anonymisées.
- 11 Après correction, le LORIA a procédé à l'alignement texte/son automatique à l'aide des outils LASTAS et JTrans¹. Le logiciel JTrans prend en entrée des fichiers au format trs ou TextGrid (Praat, Boersma & Weenink 2015). Les fichiers peuvent comporter un alignement sommaire préalable, mais ils peuvent également ne pas être alignés du tout avant traitement. L'alignement automatique se fait au mot et au phonème. Les principales difficultés d'alignement se situent au niveau des chevauchements de parole. En sortie, le logiciel génère plusieurs fichiers TextGrid, dont l'un est utilisé pour l'étape ultérieure, à savoir celle de l'anonymisation.
- 12 L'anonymisation des fichiers son s'est faite à partir de l'alignement automatique. Un fichier au format TextGrid a été généré comportant deux tiers (tierces ?) : un tiers (une tierce ?) comportant exclusivement le mot-clef « buzz », en lieu et place de chaque portion à anonymiser, et un second tiers comportant les parties à anonymiser lisibles (afin de faciliter la phase de vérification de l'alignement). Une réécoute systématique a permis de vérifier que les parties discursives en question correspondaient bien au signal sonore. Dans le cas contraire, un réalignement manuel a été effectué. Les fichiers vérifiés ont été ensuite traités par le script d'anonymisation élaboré par D. Hirst sous le logiciel Praat.
- 13 Les transcriptions au format TextGrid alignées texte/ son et anonymisées sont mises ensuite dans un format TEI en vue de la réalisation des annotations automatiques ultérieures, de la génération des divers formats de diffusion et de l'exploitation à l'aide de la plate-forme d'interrogation.
- 14 Les métadonnées comme les transcriptions étaient très hétérogènes tant au niveau des formats que du contenu. Afin de conserver l'intégralité des informations fournies tout en rendant possible les interrogations sur l'ensemble de la base, deux niveaux ont été distingués. Le premier, purement informatif, représente la transposition des métadonnées initiales, propres à chaque corpus. Elles seront mises à disposition des utilisateurs mais ne permettront pas de construire des requêtes pour constituer des sous-corpus. Le second niveau, propre à ORFÉO, est homogène et commun à l'ensemble de la base et servira à la création de requêtes. L'homogénéisation des métadonnées a nécessité un travail de saisie manuelle, seul garant de la cohérence des dénominations. Cette saisie a été faite à partir de fichiers xml générés automatiquement. Les métadonnées communes ont été corrigées et complétées dans l'éditeur Oxygen, à l'aide d'un schéma de document contraignant la saisie afin de limiter les erreurs.
- 15 Les principales métadonnées communes que nous avons retenues sont les suivantes : identifiant du corpus, type (oral/écrit), langue, identifiant du fichier, responsable(s), nature (entretien, réunion, transaction, etc.), milieu (professionnel, privé), médium (face-à-face, téléphone, radio, etc.), résumé, date, durée, qualité du son, lieu, nombre de locuteurs, degré de planification (non-planifié, semi-planifié, planifié), conditions

d'utilisation, nombre de mots, identifiant du locuteur, âge, sexe, statut de la langue française, profession, niveau d'études, lieu de naissance. Ces métadonnées seront ensuite validées par les producteurs des ressources puis projetées dans des en-têtes TEI.

- 16 Ce travail a été possible grâce à l'intervention d'un grand nombre de personnes. En premier lieu, les producteurs des corpus mais aussi les étudiants, qui ont été rémunérés pour réaliser les tâches de corrections, de saisie et d'anonymisation, des membres du laboratoire ATILF (B. Husson, B. Gaiffe, J. Perignon) qui sont intervenus à divers stades et ont assuré, notamment, les aspects informatiques, L. Bérard qui a suivi de près l'ensemble des étapes et a assuré de nombreuses vérifications automatiques et manuelles, des membres du LORIA (C. Cerisara, D. Fohr, O. Mella, D. Jouvét) pour l'alignement automatique, L. Lampen pour la mise en place du système de versionnage et C. Étienne pour le travail de réflexion sur les métadonnées. C. Benzitoun a assuré la coordination de l'ensemble.
- 17 Nous avons également bénéficié des discussions et des avancées du consortium corpus oraux (IRCOM, France) et du groupe ISO (Europe) pour la TEI. Une table ronde sur les formats de transcription et les métadonnées a été organisée, en juin 2014, par C. Benzitoun, O. Baude, C. Étienne et C. Parisse. Un atelier sur les métadonnées dans les corpus oraux a été présenté dans le cadre de la journée de bilan IRCOM, en septembre 2014, par C. Benzitoun et C. Étienne.
- 18 Le tableau ci-dessous synthétise les données qui ont été traitées et seront mises à disposition sur la plate-forme Ortolang. Un tableau détaillé sera mis à disposition sur le site du projet². Il s'agit pour l'instant d'une approximation, les données qui seront diffusées pouvant être légèrement différentes de ce qui est mentionné.

Figure 1. Corpus oraux mis à disposition en fin de projet

Identifiant du corpus	Nb. de fichiers	Taille (mots)	Durée
Corpus d'entretiens	10	13 000	1h
Corpus de référence du français parlé	134	440 000	37h
Corpus de français parlé parisien	34	500 000	40h
C-ORAL-ROM	175	300 000	25h
Corpus d'entretiens (Y. Kawaguchi)	77	728 000	45h
Corpus d'entretiens	37	62 000	5h
Corpus Clapi	14	210 000	16h
Corpus domaine académique (Fleurion)	51	40 000	3h
Corpus réunions	29	200 000	24h

Corpus contes (French Oral Narrative)	87	140 000	16h
Corpus VALIBEL	74	450 000	40h
Corpus TCOF	98	400 000	35h
Corpus OFROM	122	330 000	28h
TOTAL	942	3 813 000	315h

5. Le système d'annotation morphosyntaxique ORFÉO

- 19 Le système d'annotation syntaxique doit permettre de faire des requêtes sur les textes du corpus enrichis d'analyses en parties de discours et en relations de dépendances. La ressource majeure pour l'entraînement des outils est le French Treebank (Abeillé *et al.*, 2003), conçu à partir de textes écrits. Il n'existe pas de corpus de taille suffisante pour le français oral. Le corpus annoté Rhapsodie (Lacheret-Dujour, Kahane & Pietrandrea, à paraître) ne comporte que 30 000 mots. Le corpus DECODA, que nous utilisons comme ressource auxiliaire, comporte certes 800 000 mots partiellement annotés à la main, mais le caractère particulier de son origine (conversations téléphoniques) ne permet pas de garantir les performances de l'analyseur sur d'autres types de corpus (en particulier de longs monologues). Pour la partie écrite, les performances de notre analyseur sont comparables à celles des autres ressources : 86 % d'analyses correctes en moyenne. Des aménagements ont été nécessaires pour obtenir des résultats comparables sur l'oral. Il reste qu'il convient d'améliorer cette performance, si l'on veut que la ressource serve d'outil à des analyses linguistiques. Une des retombées du projet pourrait être d'ailleurs d'évaluer quel est le niveau de performance des annotations automatiques qui est requis pour rassembler des données fiables en vue d'une analyse linguistique. Les études pilotes prévues dans le projet nous permettront d'évaluer notre outil dans ce sens. Le système d'annotation prévu cherche avant tout à être fiable. Cette recherche de fiabilité peut amener à des analyses qui ne sont pas immédiatement utilisables pour des recherches linguistiques. Cet inconvénient sera en partie supprimé grâce aux corrections automatiques des sorties de l'analyseur pour rapprocher l'annotation des standards de l'analyse linguistique. Il faut enfin signaler que la précision de l'annotation automatique ou semi-automatique n'est qu'un élément de l'efficacité d'un corpus outillé. L'autre est la performance et la facilité d'utilisation du langage de requêtes qui permet de rassembler toutes les configurations répondant à un certain schéma d'annotation. Le projet ORFÉO comporte la construction d'un tel langage. Bien qu'interdépendants dans la pratique, ces deux outils ne relèvent pas des mêmes compétences informatiques. En tout état de cause, si de nombreux corpus oraux ont été automatiquement annotés en parties du discours (PoS), l'annotation automatique en relations de dépendance conçue par le LIF peut être considérée comme pionnière pour le domaine français.
- 20 Nous commencerons par recenser les difficultés particulières posées à l'annotation automatique par la spécificité des corpus oraux, en indiquant les stratégies mises en

œuvre pour les surmonter. Dans une deuxième partie, nous exposerons les grandes lignes des outils en construction pour l'outillage ORFÉO.

5.1 Les difficultés rencontrées dans l'annotation morphosyntaxique de l'oral spontané

5.1.1 La segmentation

- 21 Les corpus oraux que nous rassemblons ne comportent pas les marques de ponctuation qui servent de balises pour fixer le domaine d'action des analyseurs à l'écrit. L'analyseur ne peut fonctionner efficacement que sur des segments de taille limitée. Pour le corpus DECODA, constitué de demandes de renseignement téléphoniques, la segmentation en tours de parole suffisait, mais il a fallu créer des balises pour les autres corpus et donc réviser la segmentation de DECODA. Nos tentatives de créer un système de segmentation automatique en unités de texte correspondant aux phrases de l'écrit n'ayant, pour l'instant, pas débouché sur la détermination de balises fiables, nous avons décidé de segmenter à la main 1 000 000 de mots du corpus oral pour tester les outils le plus vite possible. Un guide de segmentation a été écrit pour assurer le maximum de convergence entre les annotateurs. Mais il faut accepter un certain degré de variation. Variation que l'on observe d'ailleurs dans la ponctuation des corpus écrits, et qui ne sera pas normalisée, par exemple dans l'utilisation du point, balise utilisée par les parseurs disponibles (Deulofeu, 2011).

5.1.2 La tokenisation

- 22 Une source d'erreurs importante tient au regroupement d'ordinaire effectué dès le dictionnaire de plusieurs mots en un seul composé. On relève ainsi dans le dictionnaire LEFF, qui avait été utilisé tel quel pour le traitement du corpus DECODA, la conjonction composée bien#que. Cette prise de position *a priori* interdit, de fait, une analyse correcte des séquences où les deux termes du composé doivent être dissociés : par exemple dans *je sais bien que tu voulais venir*. Pour pallier cette difficulté, nous avons décidé d'organiser de façon originale la chaîne de traitement : il s'agit de retarder la détermination des composés de ce type le plus possible, c'est-à-dire au moment du passage de l'analyseur syntaxique. La composition est considérée comme une relation syntaxique spécifique (dépendance MORPH) que l'analyseur doit apprendre à placer. Les premiers essais dans ce domaine sur diverses conjonctions au départ composées dans le dictionnaire donnent de bons résultats. La démarche sera donc étendue à d'autres cas (déterminants, adverbess).

5.1.3 Analyse syntaxique

- 23 Il faut distinguer la question des constructions caractéristiques de l'oral spontané et les obstacles que la « performance » met à la bonne reconnaissance des constructions quelles qu'elles soient. Comme exemple du premier cas, on peut prendre la question des périphériques.

5.1.3.1 Traitement des périphériques

- 24 Dans l'exemple qui suit, on a un groupe nominal, *les chaussures*, qui ne peut être analysé de façon satisfaisante par aucune des relations existantes dans le jeu des relations syntaxiques standard.
- (4) toi les chaussures tu aurais jamais dû essayer de faire des économies
- 25 Nous allons mettre au point une procédure d'évaluation pour choisir entre deux traitements possibles de cette structure :
- affecter le 'hanging topic' *les chaussures* de la même relation de dépendance (ajout) que les cas canoniques d'ajout non prépositionnel : *la semaine dernière* ;
  - introduire un nouveau lien PERIPH pour les cas de 'hanging topic'.
- 26 On rencontre un problème voisin avec des périphériques phrastiques. Ainsi, dans l'exemple suivant, prononcé par un agriculteur qui insiste sur le travail préparatoire à sa participation au Salon de l'agriculture,
- (5) on part pas à Paris on prend la vache et on y va //
- 27 on aimerait analyser la séquence comme une parataxe à l'intérieur d'une seule unité de segmentation. Ce qui permettrait d'obtenir directement la bonne interprétation : « on part pas à Paris en se contentant de prendre la vache et d'y aller ». Cette interprétation est impossible si l'on segmente *on part pas à Paris* comme une unité indépendante. L'analyse syntaxique souhaitable impliquerait donc qu'on établisse une relation de dépendance (PARAtaxe) entre les deux verbes, ce qui n'est pas canonique mais qui permettrait d'analyser la séquence en une seule unité. L'analyse pourrait être étendue aux exemples :
- (6) il y a trois semaines il est venu
  - (7) elle s'est mariée elle avait trois ans
  - (8) il y a des gens ils n'ont pas de quoi vivre
- 28 Mais là encore le choix d'une solution implique de tester la capacité pour le programme de reconnaître avec fiabilité les liens modélisant la parataxe. En cas d'échec, on adoptera une analyse sous-spécifiée.

5.1.3.2 Les obstacles aux bons rattachements dans les constructions canoniques

- 29 Il s'agit de phénomènes liés aux modes de production des énoncés propres à l'oral spontané. Leur point commun est d'introduire des discontinuités dans la régularité des séquences syntaxiques canoniques, soit en créant des séquences agrammaticales (disfluences), soit en éloignant le dépendant de son gouverneur.

5.1.3.2.1 Disfluences

- 30 Sous ce terme classique, on regroupe des phénomènes qui ne sont pas sans intérêt pour l'analyse linguistique, notamment pour mieux comprendre les conditions d'insertion lexicale dans les positions syntaxiques. Il ne s'agit donc pas d'en perdre la trace. Pour ce faire, nous avons choisi de laisser le programme d'analyse syntaxique décider lui-même du statut de disfluence. La stratégie consiste à rattacher dans le corpus d'apprentissage la partie disfluente à la séquence régulière par un lien arbitraire DISLINK. Comme on le voit dans la figure 5 (ci-dessous) où l'amorce *dans le* de « *enregistré dans le dans le métro* » est rattachée à *enregistré* sans que cela perturbe le bon rattachement du dépendant non disfluent. L'analyseur syntaxique apprend donc à placer le lien DISLINK. Le dispositif

fonctionne bien pour des répétitions de segments courts, qui peuvent être repérés automatiquement. Nous cherchons à l'étendre à d'autres cas, ainsi qu'à donner une représentation moins arbitraire des disfluences. Par exemple en établissant une relation de type coordination entre séquence disfluente et séquence régulière.

5.1.3.2.2 Parenthèses

- 31 Après une parenthèse, on constate souvent qu'un constituant soit n'est pas rattaché du tout, soit est rattaché à un mauvais gouverneur interne à la parenthèse. Ainsi dans l'exemple suivant **qu'** gouverné par *a (conscience)* est rattaché à tort à *ressens* :

(9) et là on a conscience depuis quelques mois (enfin c'est ce que je ressens) **qu'**il faudra encore peut-être bien une génération

- 32 Pour pallier cette difficulté, l'équipe a décidé d'annoter les parenthèses dans le corpus d'entraînement (CE). Leur contenu sera ignoré au cours de l'analyse syntaxique. La solution peut être étendue aux incises (*je crois, dit-il*) et de façon générale à tous les éléments y compris certains types de disfluences qui viennent s'insérer entre un gouverneur et un dépendant sans être eux-mêmes intégrés à la structure grammaticale de la phrase.

5.1.3.2.3 Énumérations / listes

- 33 Les énumérations sans coordonnants sont difficiles à analyser, car on ne peut s'appuyer sur un coordonnant pour commander les relations de coordination et, par là même, les rattacher au contexte :

(10) ça on n'en veut pas des exécutants et puis qui en même temps sont des pompiers inefficaces et impuissants devant les conséquences de cette politique / la misère / le chômage / la délinquance

- 34 Nous avons choisi d'annoter manuellement ces cas par un slash (l'équivalent de la virgule de l'écrit) devant chaque terme de l'énumération, ce qui permet une analyse standard en coordination.

5.2 Organisation de l'annotation automatique

5.2.1 La chaîne de traitement

- 35 À partir des considérations précédentes, nous avons adopté la chaîne de traitement suivante pour l'oral :

- Segmentation
- Tokenisation (avec un minimum de mots composés)
- Étiquetage en PoS
- Lemmatisation
- Intégration des informations du dictionnaire de valences³
- Élimination des « parenthèses »
- Analyse en dépendances

- 36 Le dispositif est tel que l'analyseur a accès à une partie de l'information contenue dans Dicovalence (caractéristiques syntaxiques du complément, réalisations possibles en POS, traits sémantiques des compléments (+/- subjonctif) pour désambiguïser certaines séquences.

5.2.2 Analyse morphosyntaxique

5.2.2.1 Méthode

- 37 Elle consiste à entraîner un programme probabiliste sur un corpus oral segmenté, analysé manuellement en POS et Dépendances (désigné par CE). L'évaluation est faite sur un corpus distinct annoté manuellement (désigné par CT). Le CE choisi pour l'oral est le corpus DECODA. Il est constitué de conversations téléphoniques finalisées. Les tours de parole généralement assez brefs sont utilisés comme segmentation. Une partie du corpus sera re-segmentée à la main.

5.2.2.2 Les jeux d'étiquettes

5.2.2.2.1 Critères de choix

- 38 Nous avons adapté le jeu utilisé pour DECODA. Plusieurs critères ont guidé le choix des étiquettes. Nous montrons sur quelques exemples comment ils ont été mis en œuvre et articulés.

5.2.2.2.2 Fiabilité des résultats

- 39 L'analyse de l'item *de* comme PRE (préposition) ou DET dans : *je n'ai pas de feu* est l'objet de nombreuses controverses linguistiques. Et son analyse automatique, source de nombreuses erreurs. Nous avons testé deux solutions. Soit distinguer dans le CE l'analyse en PRE et en DET, comme pour DECODA ; soit affecter à *de* la seule catégorie PRE. La reconnaissance étant meilleure dans le deuxième cas, nous avons décidé de toujours l'analyser comme une PRE, en laissant à l'analyse syntaxique la tâche de distinguer les différentes constructions où entre cette PRE.
- 40 Nous voulions cependant éviter que ce choix ait des conséquences indésirables dans l'analyse en relations de dépendance. Avec le jeu de relations original, (11) et (12) auraient eu la même analyse en dépendance (complément prépositionnel), et (12) et (13), des analyses différentes, ce qui est contre-intuitif sur le plan linguistique.
- (11) je mange des amandes
(12) je parle des amandes
(13) je grille les amandes
- 41 C'est la raison pour laquelle nous avons introduit la relation SPE (spécifieur). La préposition peut alors être analysée comme spécifieur dans (11), ce qui permet de dire que le complément *y* est « direct » comme en (13), tandis qu'elle est complément du verbe dans (12), qui présente donc un complément indirect.

5.2.2.2.3 Adéquation avec analyse linguistique admise

- 42 Comme les autres jeux de catégories et de relations, celui de DECODA reprend assez largement les catégories de la grammaire scolaire. Cette démarche est cohérente avec l'idée que l'analyseur ne fournit pas une analyse définitive des structures, mais permet de rassembler des données servant de base à des analyses qui peuvent être d'orientations théoriques différentes. La grammaire implicite sous-jacente à l'établissement des étiquettes doit donc tendre vers cette *Basic Linguistic Theory*, dont le descriptiviste Dixon dit qu'elle est la grammaire traditionnelle, moins les erreurs les

plus flagrantes. C'est, en fait, cette direction que nous avons suivie en intégrant à notre grammaire implicite certains acquis de la linguistique descriptive. La différence principale tient à l'articulation catégorie/fonction. La grammaire traditionnelle distingue plusieurs fonctions (épithète, complément de nom) en relation avec les catégories concernées (adjectif, groupe nominal), là où les approches descriptives ne voient qu'une seule relation (complément) pouvant porter sur plusieurs catégories.

- 43 Ce principe nous a amenés à réduire le nombre de relations par rapport au jeu d'étiquettes DECODA, comme on le verra plus bas ; à supprimer la catégorisation multiple en adverbe, préposition, conjonction, en fonction du contexte pour des mots comme *pour*, *sans*, *après*, au profit de la classification unique comme PRE, adoptée par la linguistique descriptive ; à distinguer la catégorie déterminant (DET) de la relation syntaxique Spécifieur (SPE). Le SPE est le dépendant gauche du N en position sujet. Les DET sont les items qui ne peuvent exercer que cette fonction. Le dépendant SPE peut-être un DET (**le** livre) un adjectif (**quelques** livres), une préposition : j'ai **de** beaux livres... Parfois, le critère d'adéquation descriptive entre en conflit :
- tantôt avec celui de fiabilité des résultats produits par l'analyseur.
- 44 Ainsi, la catégorie PRQ (pronom *qu-*) est en adéquation avec les descriptions contemporaines en ce qu'elle ne distingue pas pronoms relatifs et pronoms interrogatifs au niveau des POS et efface la distinction entre pronoms (*quoi*) et adverbes (*où*) relatifs. Mais nous avons maintenu, malgré son inadéquation descriptive, la double catégorisation traditionnelle de *que* en PRQ et CSU (conjonction de subordination) et classé *qui* sujet PRQ pour faciliter la reconnaissance des relatives par rapport aux complétives. En effet, les programmes ne pouvant reconnaître ce qui serait dans certaines analyses linguistiques une catégorie vide, analyser les *que* dans les relatives comme conjonctions (complémentisateurs) aboutirait à donner la même analyse syntaxique à *le fait qu'il a signalé* et *le fait qu'il ait regretté ça*.
- tantôt avec celui de la facilité de compréhension par l'utilisateur.
- 45 En dehors des classes réduites CLI (clitique) et PRQ, on maintient la distinction entre pronom (*lui*, *quelqu'un*) et adverbe (*là*, *quelque part*) sur critère morphologique conformément aux analyses les plus courantes.

5.2.2.3 Jeu d'étiquettes POS

- 46 L'adaptation du jeu DECODA s'est faite de façon automatique. Le travail principal a été la révision du dictionnaire LEFFF, en particulier pour le traitement des composés. Le tableau ci-dessous présente les étiquettes retenues dans le projet.

Figure 2. Liste des étiquettes retenues

Tag	Catégorie
ADJ	Adjectifs
ADN	Adverbes négatifs
ADV	Adverbes
CLI	Clitiques sujets / objets / réfléchis
CLN	Clitiques négatifs
COO	Conjonctions de coordination
CSU	Conjonctions de subordination
DET	Déterminants
ETR	Mots étrangers
INT	Interjections
NOM	Noms
PCT	Signes de ponctuation
PRE	Prépositions
PRF	Préfixes
PRO	Pronoms indéfinis / démonstratifs / possessifs / toniques
PRQ	Pronoms relatifs / interrogatifs
TRC	Troncations
VNF	Verbes à l'infinitif
VPP	Participes passés
VPR	Participes présents
VRB	Verbes conjugués

- 47 La classe interjection regroupe les mots qui ne peuvent avoir de gouverneurs (sauf verbes introducteurs de discours direct), à savoir les interjections classiques (*eh, hélas, bof...*) et des items qui sont aujourd'hui classés comme particules discursives (*euh, ben, voilà...*). La performance globale de l'étiqueteur en POS est de 98,25 % conforme aux standards en matière d'analyse de l'oral. L'évaluation a été faite sur une partie de DECODA annotée manuellement.

5.2.2.4 Jeu d'étiquettes des relations de dépendances

- 48 Nous sommes partis du jeu de relations utilisé pour DECODA, soit :

Figure 3. Relations utilisées pour le corpus DECODA

OBJ	objet direct
AFF	clitique sans fonction (<i>s'en aller</i>)
D-COORD	dépendant de la conjonction de coordination
MOD	modifieur (complément autre qu'objet du verbe)
SUJ	sujet

ROOT	racine de l'énoncé ou absence de gouverneur
MOD_REL	lien antécédent relative
AUX	lien verbe-auxiliaire
P_OBJ	objet prépositionnel
COORD	dépendant de la conjonction de coordination
DET	déterminant

- 49 Nous l'avons modifié à la suite de l'évaluation des résultats qui peut être résumée dans le tableau ci-dessous. Le tableau indique d'abord la relation, puis sa fréquence, puis la précision de la reconnaissance, et enfin l'impact de la précision en fonction de la fréquence.

Figure 4. Evaluation des résultats

LABEL	FREQ	ACC	IMPACT
OBJ	19.55	90.51	11.32
AFF	1.20	94.08	0.44
D-COORD	2.31	83.56	2.32
MOD	12.86	75.16	19.52
SUJ	11.37	95.89	2.86
ROOT	26.50	80.83	31.06
MOD_REL	0.65	65.24	1.38
AUX	1.53	98.71	0.12
P_OBJ	4.02	71.30	7.05
COORD	1.41	69.58	2.62
DET	8.39	94.19	2.98
DISFLINK	10.19	70.48	18.41

- 50 En général, on observe 87 % de bons gouverneurs et 84 % de bonnes étiquettes. (Le corpus d'entraînement et de test est DECODA). Le rappel des résultats pour l'écrit de référence (French Treebank : 87 %).
- 51 La Figure 5 présente le résultat d'une analyse correcte sous forme tabulaire simplifiée. La première colonne donne l'ordre linéaire des items ; la deuxième, l'item ; la troisième,

le lemme ; la quatrième, la POS ; la cinquième, la position du gouverneur ; et la dernière, la fonction de l'item numéroté dans la première colonne. Cette présentation sera transformée en un format plus classique de Treebank pour l'utilisateur.

Figure 5. Analyse syntaxique d'une portion de DECODA

1	ah	ah	INT	0	ROOT
2	non	non	INT	0	ROOT
3	ça	ça	CLI	7	SUJ
4	a	avoir	VRB	6	AUX
5	pas	pas	ADN	7	MOD
6	été	être	VPP	7	AUX
7	enregistré	enregistrer	VPP	0	ROOT
8	dans	dans	PRE	7	DISFLINK
9	le	le_	DET	8	DISFLINK
10	dans	dans	PRE	7	MOD
11	le	le	DET	12	DET
12	métro	métro	NOM	10	COMP

52 Ces tableaux permettent notamment de repérer les sources d'erreurs les plus fréquentes. Mais une première source d'erreurs pourrait bien découler d'un fait général : il est difficile d'annoter automatiquement des données très éloignées de celles qui ont été utilisées pour entraîner l'analyseur syntaxique. Des essais sont en cours pour pallier cette difficulté :

- Ajouter au Corpus d'entraînement d'autres types de corpus annotés en relations : Rhapsodie, le French Treebank ;
- Annoter manuellement un ensemble de « genres » différents pour ré-entraîner le parseur, voire procéder à différents paramétrages.

53 Une deuxième source d'erreurs tient sans doute au nombre de relations à identifier. On trouve ainsi, dans DECODA, pour les dépendances verbales : Pobj ; obj ; mod ; mod_rel. Leur taux de reconnaissance est très variable : 90 % pour obj, mais mod, Pobj et mod_rel ne sont reconnues qu'à moins de 75 %. Nous avons choisi d'éliminer le plus possible les relations à problèmes. Les étiquettes mod, mod_rel, Pobj sont un héritage de la tradition, mais elles représentent en fait une seule relation de dépendance au verbe, les distinctions étant en réalité redondantes avec l'analyse en POS des compléments (NP, PP, ADV...). Nous avons donc réduit toutes ces relations à deux : OBJ pour objet direct et COMP (pour les autres cas). L'utilisation des informations contenues dans le dictionnaire de valences permet ensuite de distinguer sur des bases fiables, parmi les COMP, ceux qui appartiennent à la valence des verbes, seule distinction à la fois

linguistiquement pertinente et utile pour la désambiguïsation des séquences. Nous travaillons donc, pour l'instant, avec le jeu suivant de relations, présentées ici en fonction du gouverneur :

- Pour le gouverneur Verbe, on retient les relations de dépendance AUX, COMP et OBJ.
- Pour les autres gouverneurs (ADJ, NOM, ADV, PRE...), on utilise SPE et COMP.
- Les conjonctions de coordination ont un système particulier de dépendances, elles sont reliées par la relation COOR à leur gouverneur et par la relation COMP à leur dépendant.

- 54 Des évaluations du nouveau dispositif sont en cours sur une partie segmentée du corpus ORFÉO. Nous constituerons à partir de là une liste de constructions présentant des difficultés d'analyse et donc exigeant un approfondissement du traitement syntaxique.

6. Plate-forme d'accès et de requêtes

- 55 Le travail est en cours, notamment grâce à Lari Lampen, en contrat d'ingénieur d'études, sous la responsabilité de Kim Gerdes et de Sylvain Kahane. La première tâche a consisté à évaluer les plates-formes actuelles d'interrogation de corpus arborés (treebanks). L'outil Annis, développé actuellement à l'Université de Postdam (Allemagne), a été jugé le plus performant pour le projet, sous réserve de modifications. La première concerne la possibilité d'accepter plusieurs formats qui peuvent être appareillés : TEI, CONLL.
- 56 La deuxième modification en cours doit permettre de gérer des fichiers à plusieurs stades de développement, avant ou après intégration de certaines annotations, lesquelles peuvent ensuite être modifiées. De même, les formats d'encodage peuvent être amenés à évoluer, notamment lors de l'intégration à la plate-forme Ortolang dans la phase finale.
- 57 L'interface utilisateur prévoit, outre la constitution d'un corpus de travail à partir des métadonnées des corpus existants, plusieurs niveaux d'interrogation. Un niveau de requête simple aboutira à une consultation « à la Google » et à un tableur présentant des indications statistiques sur la fréquence de l'item choisi. Un niveau intermédiaire sera dédié à la recherche par lemme et POS. Ces deux interfaces constituent une surcouche par rapport à l'interface de requêtes d'Annis, qui permet, elle, des interrogations en dépendances.
- 58 Par ailleurs, les données seront interrogeables directement par un lien vers les différents corpus sources de la plate-forme. Le principe est que cet outil soit le plus aisé possible à manipuler. Des fenêtres d'aide permettront aux utilisateurs novices de comprendre le fonctionnement des outils mis à disposition (concordancier et requêtes complexes). L'objectif du projet est en effet d'encourager l'utilisation des corpus et des outils, que ce soit à des fins de recherche ou d'enseignement. Au moment de la rédaction de cet article, le projet vient de se voir accorder une prolongation. La plate-forme et les données qui lui sont associées seront donc disponibles courant 2017.

BIBLIOGRAPHIE

- Abeillé A., Clément L. & Toussenet F. (2003). « Building a treebank for french », in A. Abeillé (éd.) *Treebanks*. Kluwer : Dordrecht.
- Bawden R., Botalla M.-A., Gerdes K. & Kahane S. (2014). « Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie », *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, Reykjavik.
- Benzitoun C. & Bérard L. (2010). « Mutualisation et uniformisation de ressources de français parlé », *Cahiers de praxématique* 54-55 : 175-188.
- Bérard L. (2014). « Dépendances à longue distance et genres textuels », *Actes du Congrès mondial de linguistique française*, 2349-2365.
- Boersma P. & Weenink D. (2015). *Praat : doing phonetics by computer [Computer program]*. Version 5.4.18, retrieved 7 September 2015 from <http://www.praat.org/>.
- Botalla M.-A., (2014). *Analyse du flux de dépendance dans un corpus de français oral annoté en microsyntaxe*, Mémoire de master, Université Sorbonne Nouvelle.
- Deulofeu J. (2011). « Peut-on établir un système de ponctuation des transcriptions de textes oraux linguistiquement fondé », *Langue française* 171.
- Groupe ICOR (Bruxelles S., Jouin-Chardon E., Traverso V.) & Guinamard I. « “Du coup” dans l’interaction orale en français : description de ses usages situés à partir d’une base de données multimédia, et considérations didactiques », in *Synergie pays riverains du Mékong*.
- Lacheret-Dujour A., Kahane S. & Pietrandrea P. (en préparation). *Rhapsodie : a Prosodic and Syntactic Treebank for Spoken French*. Amsterdam : Benjamins.
- Nasr A., Bechet F., Favre B., Bazillon T., Deulofeu J. & Valli A. « Automatically Enriching Spoken Corpora with Syntactic Information for Linguistic Studies », in *International Conference on Language Resources and Evaluation (LREC)*, mai 2014.
- Wang I. (2013). *Segmentation automatique d’un corpus de français oral en unités macrosyntaxiques*, Mémoire de master, Université Sorbonne Nouvelle.
- Wang I., Kahane S. & Tellier I. (2014). « Macrosyntactic Segmenters of a spoken French Corpus », *9th Language Resources and Evaluation Conference (LREC)*, Iceland, 1-6.
- Liste des corpus oraux constitués et mis à disposition partiellement ou intégralement par les ayants droit :**
- French Oral Narrative – Janice Carruthers
<http://frenchoralnarrative.qub.ac.uk/>
- Corpus du français parisien des années 2000 – Sonia Branca, Serge Fleury, Florence Lefevre
<http://cfpp2000.univ-paris3.fr/>
- Corpus oral de français parlé en Suisse romande – Mathieu Avanzi, Marie-José Béguelin et Federica Diémoz
<http://www11.unine.ch/>
- Corpus de référence du français parlé – Équipe Delic
<http://sites.univ-provence.fr/~veronis/pdf/2004-presentation-crfp.pdf>

C-ORAL-ROM – E. Cresti, M. Moneglia

<http://lablita.dit.unifi.it/coralrom/>

Corpus de langue parlée en interaction – V. Traverso, L. Mondada, S. Bruxelles, C. Étienne,

E. Jouin-Chardon, S. Teston Bernard

<http://clapi.ish-lyon.cnrs.fr/>

Corpus VALIBEL – Discours et variation – Anne Catherine Simon

<http://www.uclouvain.be/valibel>

Corpus TUFS – Yuji Kawaguchi

[http://www.coelang.tufs.ac.jp/multilingual_corpus/fr/index.html?](http://www.coelang.tufs.ac.jp/multilingual_corpus/fr/index.html?contents_xml=corpus&menulang=en)

[contents_xml=corpus&menulang=en](http://www.coelang.tufs.ac.jp/multilingual_corpus/fr/index.html?contents_xml=corpus&menulang=en)

Corpus Traitement de corpus écrits et oraux (TCOF) – V. André, C. Benzitoun, E. Canut, J.-

M. Debaisieux

<http://www.cnrtl.fr/corpus/tcof/>

Corpus Français langue étrangère universitaire, Ressources et outils numériques (FLEURON) –

V. André, M. Ciekanski, F. Poncet, J.-M. Debaisieux

<https://apps.ATILF.fr/fleuron2/>

(site en cours d'expérimentation)

Corpus d'entretiens – Ensemble de transcriptions fournies par S. Caddéo, J.-M. Debaisieux, et

élaborées dans le cadre d'enseignements

Corpus de réunions de travail – M. Husianycia (élaboré dans le cadre de son travail de thèse)

http://www.ATILF.fr/IMG/pdf/theses/These_Husianycia_Magali_2011.pdf

Corpus Rhapsodie – A. Lacheret

<http://www.projet-rhapsodie.fr/>

NOTES

1. Le logiciel JTrans est librement téléchargeable à l'adresse suivante : <https://github.com/synalp/jtrans/releases/latest>.
2. <http://www.projet-orfeo.fr/>
3. Le dictionnaire choisi est Dicovalence (www.bach.arts.kuleuven.be/dicovalence/).

RÉSUMÉS

L'article présente le projet ORFEO (outils pour l'étude du Français écrit et oral). Ce projet a consisté à rassembler 4M de mots de Français oral et 6M de mots d'écrit à partir de ressources existantes. Les ressources orales ont été unifiées dans un format trs et alignées avec les fichiers sons au phonème. Les corpus sont présentés avec des métadonnées qui permettent de constituer des sous-corpus d'étude. Les données orales ont été segmentées et annotées en POS et relations de dépendance semi automatiquement. La précision des annotations a fait l'objet d'une

évaluation. L'ensemble est interrogeable à partir d'une plateforme adaptée du logiciel libre ANNIS. La ressource sera mise à disposition en open source courant 2017.

The paper presents the outcomes of the project ORFEO (tools for the study of spoken and written French). The outcomes consist of a 4M words spoken French and 6M written French sampled corpus. Detailed metadata allows the user to build his own study corpus according to his research purposes. Oral resources have been unified under trs. Format aligned with sound files. Spoken data have been segmented and annotated into POS and dependency relations. The precision of these annotations has been evaluated and meets the current standards. Queries are possible through a platform adapted from ANNIS free software. The whole resource will be available during the year 2017.

INDEX

Keywords : equipped corpus, spoken and written French, POS annotation, dependency annotation

Mots-clés : corpus outillé, français parlé, français écrit

AUTEURS

CHRISTOPHE BENZITOUN

Université de Lorraine, Atilf UMR 7118

JEANNE-MARIE DEBAISIEUX

Université Paris 3 Sorbonne Nouvelle, Lattice UMR 8094

HENRI-JOSÉ DEULOFEU

Université Aix-Marseille, Lif UMR 7279