TALN 2009, Senlis, 24–26 juin 2009

Analyse syntaxique en dépendances de l'oral spontané

Alexis Nasr[1] Frédéric Béchet[2]
(1) LIF - CNRS - Université Aix Marseille
(2) LIA - Université d'Avignon

Résumé. Cet article décrit un modèle d'analyse syntaxique de l'oral spontané axé sur la reconnaissance de cadres valenciels verbaux. Le modèle d'analyse se décompose en deux étapes : une étape générique, basée sur des ressources génériques du français et une étape de réordonnancement des solutions de l'analyseur réalisé par un modèle spécifique à une application. Le modèle est évalué sur le corpus MEDIA.

Abstract. We describe in this paper a syntactic parser for spontaneous speech geared towards the identification of verbal subcategorization frames. The parser proceeds in two stages. The first stage is based on generic syntactic ressources for French. The second stage is a reranker which is specially trained for a given application. The parser is evaluated on the MEDIA corpus.

Mots-clés : Analyse syntaxique, reconnaissance automatique de la parole.

Keywords: Syntactic parsing, automatic speech recognition.

1 Introduction

De nombreux systèmes automatiques contemporains de compréhension de la parole produisent des représentations du sens des énoncés sous la forme de cadres sémantiques à la FrameNet (Baker *et al.*, 1998). La construction de ces représentations suppose en général une étape d'analyse syntaxique plus ou moins poussée. Cette dernière se situe habituellement en aval d'un système automatique de reconnaissance de la parole (SRAP), qui produit, pour un signal acoustique donné, correspondant à un énoncé, la retranscription jugée la plus probable par le SRAP ou bien les n retranscriptions les plus probables représentées sous la forme d'un graphe de formes. L'analyse syntaxique d'un tel graphe est confrontée à de nombreux problèmes, en sus des problèmes traditionnels de l'analyse syntaxique. Ceux-ci sont de natures diverses, provenant des spécificités de l'oral par rapport à l'écrit, mais aussi du caractère imparfait des retranscriptions produites par le SRAP, surtout lorsque la parole à transcrire est spontanée. Du fait de ces particularités, il est illusoire de recourir, pour une telle tâche, à des analyseurs syntaxiques probabilistes modernes, conçus pour traiter la langue écrite. Il n'est pas possible non plus de réentrainer ces analyseurs sur des corpus de l'oral, car nous ne disposons pas actuellement de corpus syntaxiques pour l'oral (où chaque enoncé est associé à un arbre d'analyse syntaxique).

Face à une telle situation, une solution consiste à développer des analyseurs ad-hoc, adaptés à un registre de langue donné, à un champ sémantique particulier, voire même à un SRAP donné.

[1]Ces travaux sont financés par l'ANR, dans le cadre du projet EPAC, contrat numéro ANR-06-MDCA-006.

Une telle approche, même si elle peut donner des résultats satisfaisants (Seneff, 1992), pêche par son manque de généricité et rend l'adaptation de tels systèmes particulièrement coûteux.

Une solution alternative consiste à distinguer, parmi les connaissances mobilisées par un tel système, ce qui relève de la syntaxe générale de ce qui est propre à un domaine applicatif. C'est dans cette direction que s'inscrit le travail présenté ici. Il propose un traitement en deux étapes.

Dans un premier temps, une analyse syntaxique partielle des sorties du SRAP est effectuée. Elle repose sur des ressources génériques, en particulier une grammaire syntagmatique partielle, ce qui n'est pas une nouveauté dans le domaine de l'analyse syntaxique automatique de l'oral (voir (Antoine *et al.*,)), mais aussi, et c'est une originalité de notre approche, sur deux ressources générales que sont le lexique syntaxique dicovalence (van den Eynde & Mertens, 2003) et sur la grammaire d'adjonction d'arbres du français FXMG (Crabbé, 2005b). Cette étape produit, pour un graphe donné de formes, plusieurs instanciations de cadres de valence. Il s'agit d'assigner aux verbes présents dans le graphe un cadre valenciel (verbe intransitif, verbe transitif direct ...) ainsi que ses actants (sujet, objet direct, objet indirect ...). Nous avons opté pour ce type d'analyse syntaxique, plutôt qu'une analyse syntagmatique plus traditionnelle, car elle fournit des représentations plus adaptées au calcul de cadres sémantiques mentionnés ci-dessus.

La seconde étape du traitement prend en entrée les cadres de valence instanciés produits à l'issue de l'étape précédente et les réordonne en fonction d'un modèle spécifique à une tâche particulière. Cette étape repose sur un modèle de boosting (Schapire & Singer, 2000). Ce dernier est entrainé sur un corpus spécifique à l'application visée.

L'organisation de l'article est la suivante : dans la section 2, nous décrivons l'analyseur syntaxique, la section 3 décrit le modèle de reclassement utilisé. La section 4 est consacrée à la partie expérimentale de ce travail et la section 5 conclut l'article.

2 Analyse syntaxique

L'analyse linguistique des sorties du SRAP se décompose, de manière classique, en une série de processus qui constituent une chaîne de traitements : l'entrée d'un processus correspond à la sortie du processus précédent. Les différents modules qui constituent notre chaîne vérifient tous un certain nombre de principes, en particulier :

1 - Les modules sont monotones, ils ajoutent de l'information linguistique à leurs entrées mais ne modifient pas les informations déjà présentes.

2 - Les modules admettent plusieurs hypothèses pondérées en entrée et produisent à leur tour plusieurs hypothèses pondérées à l'aide d'une fonction de coût propre au module. Le nombre d'hypothèses produites en sortie peut être contrôlé grâce à la pondération (on ne produit que les n hypothèses de meilleurs poids).

3 - L'ambiguïté est représentée sous la forme de graphes (graphes de formes, de mots, de catégories, de syntagmes ...) qui permettent de mettre en facteur des parties communes à plusieurs hypothèses. Ces graphes sont représentés par des transducteurs finis pondérés et la majorité des traitements correspondent à des opérations standard sur les transducteurs.

Les premiers modules de la chaîne sont standards, nous ne ferons que les évoquer pour nous attarder plus longuement sur le dernier, l'analyseur en dépendances partiel.

Le graphe de formes issu du SRAP est traité, dans un premier temps, par un module lexical qui permet de regrouper certaines formes. C'est à ce niveau que sont reconnues des unités lexicales complexes telles que les locutions (*au dessus de, à mesure que* ...) ou encore les mots composés (*pomme de terre, couvre chef* ...). Ce module repose sur le lexique des formes fléchies du français (le Le*fff*) (Sagot *et al.*, 2006). Conformément au second principe ce module permet de conserver plusieurs découpages possibles, pour les cas de chevauchement d'unités lexicales complexes. Le module produit un graphe d'unités lexicales.

Ce dernier constitue l'entrée d'un module d'étiquetage morpho-syntaxique qui produit un graphe de catégories morpho-syntaxiques, chaque catégorie étant associée à une unité lexicale. L'étiqueteur repose sur un modèle de Markov caché classique dont l'estimation des paramètres a été réalisée sur le corpus paris 7 (Abeillé *et al.*, 2003).

Le graphe de catégories est alors traité par un module d'analyse morphologique qui associe à tout couple composé d'une unité lexicale et d'une catégorie une ou plusieurs analyses morphologiques. Une analyse morphologique est composée d'un lemme et d'une série de traits morphologiques. Ce module repose, tout comme le premier, sur le Le*fff*.

La quatrième étape consiste en une analyse syntaxique partielle du graphe de catégories. Il s'agit de regrouper des séquences de catégorie au sein de syntagmes non récursifs, souvent appelés *chunks*. L'idée sous jacente est de n'effectuer que des regroupements non ambigus (tel qu'une séquence *déterminant, adjectif, nom* ou encore *auxiliaire, participe passé* ...). Il n'y a en particulier pas de rattachements prépositionnels effectués à ce niveau. Les traitements sont réalisés à l'aide d'une cascade de transducteurs, à l'image de (Abney, 1996). Les grammaires locales correspondant à chaque syntagme sont produites manuellement. Tout syntagme possède une tête qui est spécifiée dans la grammaire correspondante.

Le résultat de cette suite de traitements est un graphe de syntagmes. Il constitue l'entrée du module d'analyse de dépendances syntaxiques partiel qui est l'objet de la section suivante.

Il est important de noter que, de manière générale, le nombre d'hypothèses potentielles augmente au fur et à mesure que l'on évolue dans la chaîne de traitement. A une suite de formes peut correspondre plusieurs découpages en mots. Une séquence de mots peut correspondre à plusieurs séquences de catégories, chacune pouvant correspondre à plusieurs découpages en unités syntaxiques.

2.1 L'analyseur de dépendances syntaxiques partiel

L'objectif de ce module est de retrouver des dépendances actancielles (sujet, objet direct, objet indirect, régime d'une préposition, certains complément de noms ...) dans un graphe de sortie d'un SRAP, préalablement traité par les modules décrits ci-dessus. Nous nous intéresserons ici aux seules relations actancielles ayant pour gouverneur un verbe. La principale source d'information utilisée pour détecter l'occurrence de telles dépendances est une description des cadres valenciels (appelés aussi schémas de régime ou cadres de sous-catégorisation) des verbes.

Le module doit offrir simultanément une bonne couverture, de la souplesse et de l'efficacité.

La couverture revêt deux aspects, la couverture lexicale (le nombre de verbes pour lesquels on dispose d'une description du cadre valenciel), et la couverture syntaxique (les différentes réalisations possibles d'un cadre valenciel).

La souplesse est la capacité à retrouver des occurrences de dépendances actancielles dans des entrées bruitées. Le bruit étant constitué des particularités de la syntaxe de l'oral (disfluences, hésitations) et des erreurs commises par le système de reconnaissance de parole.

La couverture est assurée par deux ressources existantes, le lexique syntaxique Dicovalence (van den Eynde & Mertens, 2003) et la grammaire d'adjonction d'arbres FXMG (Crabbé, 2005a). La souplesse et l'efficacité sont assurées par une représentation sous spécifiée des réalisations des cadres valenciels sous la forme de transducteurs finis.

Ces aspects sont décrits successivement dans les deux sections suivantes.

Couverture lexicale et syntaxique

Dicovalence est un lexique syntaxique de verbes du français. Il recence les cadres valenciels de plus de 3700 verbes. Un cadre valenciel décrit une configuration de compléments valenciels d'un verbe (le nombre, la nature et la fonction de ses complément). La description des cadres valenciels repose sur les principes de l'approche pronominale (van den Eynde & Blanche-Benveniste, 1978) : chaque position actantielle, appelée paradigme, est décrite par le paradigme des pronoms qui peuvent l'occuper. 20 paradigmes sont distingués, parmi lesquels on trouve $P0$, $P1$ et $P2$ qui correspondent grosso modo aux sujet, objet direct et objet indirect de la grammaire traditionnelle. Comme nous l'avons précisé ci-dessus, 3700 verbes sont répertoriés, qui constituent plus de 8000 entrées, un verbe pouvant être associé à plusieurs cadres valenciels. 311 cadres valenciels distincts sont répertoriés. Un cadre valenciel est décrit de manière concise par un identifiant qui en résume les caractéristiques principales. L'identifiant `P0 P1` par exemple décrit le cadre valenciel transitif direct. Il est constitué des deux paradigmes : `P0` pour le sujet et `P1` pour l'objet direct. Une entrée lexicale associe à un lemme verbal un cadre valenciel ainsi que les pronoms des différents paradigmes qui composent le cadre valenciel.

Dicovalence offre une bonne couverture lexicale (il couvre $89,3\%$ des verbes du corpus Paris 7) mais sa couverture syntaxique est par construction limitée dans la mesure où seules les réalisations pronominales des actants sont répertoriées. Pour obtenir une meilleure couverture syntaxique, il est nécessaire d'associer à tout cadre valenciel ses réalisations non pronominales. Cette étape est réalisée à l'aide de la grammaire FXMG.

La grammaire FXMG est une grammaire d'adjonction d'arbres produite automatiquement à partir d'une méta-grammaire à l'aide du logiciel XMG (Crabbé, 2005a). Nous ne ferons pas ici de présentation des grammaires d'adjonction d'arbres, rappelons simplement qu'une telle grammaire associe à toute entrée lexicale un ensemble d'*arbres élémentaires* qui décrivent, entre autre, une réalisation possible d'un cadre valenciel de l'entrée lexicale. Les arbres élémentaires sont regroupés en *familles*, chaque famille étant composée des différentes réalisations syntaxiques d'un cadre valenciel donné. Ainsi, est associé à un verbe transitif une famille composée de l'arbre représentant la réalisation nominale du sujet et de l'objet direct (*Jean mange la pomme*), d'une cliticisation de l'objet direct (*Jean la mange*), de la diathèse passive (*la pomme est mangée par Jean*) et ainsi de suite. La grammaire ainsi produite est de taille importante, elle est composée de 7600 arbres élémentaires, regroupés en 92 familles. Les familles sont de tailles inégales. A titre d'exemple la famille des verbes transitifs est composée de 159 arbres élémentaires. Comme nous l'avons précisé ci-dessus, une telle grammaire est produite automatiquement à partir d'une description plus abstraite, appelée méta-grammaire. Tout arbre élémentaire de la grammaire est caractérisé de manière bi-univoque par un ensemble de traits qui en décrivent ses caractéristiques syntaxiques (réalisation morpho-syntaxique des actants, ordre linéaire, diathèse ...). Cette représentation permet un filtrage simple et motivé de l'ensemble

des arbres élémentaires : on peut décider de ne garder que les arbres élémentaires possédant certains traits particuliers. Chaque famille est associée à un identifiant qui exprime de manière concise le cadre valenciel associé à la famille : la famille associée aux verbes transitifs est `n0Vn1` la famille des verbes di-transitifs dont l'objet indirect est introduit par la préposition *à* est `n0Vn1an2` et ainsi de suite.

La couverture de FXMG a été évaluée sur le corpus TSNLP (Lehmann, 1996) par (Crabbé, 2005b), elle est de 75% ce qui peut être considéré comme un résultat satisfaisant étant donné les caractéristiques de ce corpus (les fréquences d'occurrence des constructions ne correspondent pas à leur fréquence d'usage dans la langue).

FXMG définit donc un ensemble d'arbres élémentaires, regroupés en familles, mais n'établit pas de lien entre les entrées lexicales et les arbres élémentaires. Il est facile de voir à cette étape de notre exposé la complémentarité de dicovalence et de FXMG. Le premier possède une bonne couverture lexicale tandis que le second possède une bonne couverture syntaxique.

Pour effectuer un lien entre dicovalence et FXMG, il convient de traduire les cadres valenciels de dicovalence en noms de familles de FXMG. Cette traduction est réalisée à l'aide de règles qui associent un couple *(paradigme, pronom)* à un fragment d'identifiant de famille. A titre d'exemple, la règle `(P0, je)` $\rightarrow$ `n0` indique que la réalisation du paradigme sujet (`P0`) sous la forme du pronom *je* se traduit par `n0` dans un identifiant de famille de FXMG. Les fragments d'identifiants de familles sont ensuite concaténés[1] pour constituer des identifiants de familles.

Représentation des schémas de valence sous la forme d'automates

Nous avons décrit dans la section précédente la manière dont les deux ressources dicovalence et FXMG avaient été reliées entre elles à l'aide de règles de correspondance. Nous disposons donc, pour chacun des 3700 verbes de dicovalence, d'une ou de plusieurs familles d'arbres élémentaires qui décrivent les différentes réalisations possibles des cadres de valence de ces verbes. Afin de retrouver une occurrence d'un cadre de valence dans les sorties de l'analyseur syntaxique partiel, chaque arbre élémentaire est représenté sous la forme d'un automate fini. On effectue ensuite l'union des automates correspondant aux arbres d'une famille donnée pour constituer l'automate de la famille.

Le processus de construction d'un automate à partir d'un arbre élémentaire est simple. Il consiste à parcourir l'arbre de manière descendante (en profondeur d'abord, de gauche à droite) et de construire pour tout nœud de substitution de l'arbre une transition étiquetée par la catégorie du nœud de substitution. Nous avons représenté dans la partie droite de la figure 1 l'automate correspondant à l'arbre de la partie gauche.

On remarquera que tout état de l'automate de la figure 1 possède une transition sur lui-même pour chaque élément de l'alphabet Σ. Ce dernier est composé des différentes étiquettes de groupes syntaxiques et de catégories morpho-syntaxiques. Ces transitions permettent aux différents actants du verbe de se trouver à une distance arbitraire de ce dernier.

On pourra remarquer que l'automate correspondant à un arbre élémentaire ne représente qu'une partie de l'information syntaxique de ce dernier, plus précisément, son nombre de nœuds de substitution, leur nature et leur ordre linéaire relatif. Ce relâchement de contraintes syntaxiques peut aboutir à des rattachements erronés. Pour attenuer cet effet, des pondérations (α et β) sont

[1] Cette opération est en fait un peu plus complexe car les indices des paradigmes dans dicovalence (le `1` de `P1`) et les indices dans les familles d'arbres (le `0` de `n0` dans `n0Vn1`) n'ont pas la même sémantique. Le premier réfère grosso modo à une fonction syntaxique tandis que le second désigne une position linéaire.

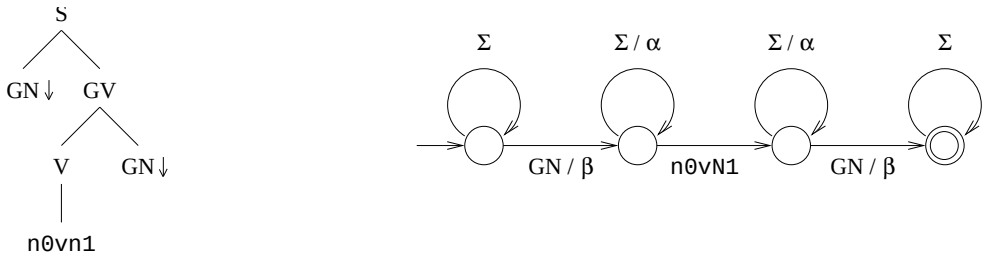

FIG. 1 – Transformation d'un arbre élémentaire en automate

ajoutées aux transitions. α matérialise une pénalité qui s'accroît au fur et à mesure que croît la distance entre le verbe et un de ses actants. β correspond à une récompense pour chaque actant trouvé. Ce jeux de pondérations permet d'implémenter une heuristique simple : on favorise la proximité entre le verbe et ses actants, ainsi que le nombre d'actants.

Le processus d'instanciation des cadres de valences dans un graphe de syntagmes issu de l'analyseur syntaxique partiel est réalisé à l'aide de l'opération de composition d'automates. De manière plus précise, nous disposons de l'automate des sytagmes, noté T_S et d'un automate par famille d'arbres élémentaires, soit n automates notés $T_{F_1}, \ldots, T_{F_n}$. Les n automates de familles sont composés successivement avec l'automate T_S et les résultats de ces compositions sont regoupés entre eux grâce à l'opération d'union[2]. Le résultat est donc $\cup_{i=1}^{n} T_S \circ T_{F_i}$[3]. Ce modèle s'oppose à une cascade de transducteurs ($T_S \circ T_{F_1} \circ \ldots \circ T_{F_n}$). Cette différence est importante dans le cas de phrases complexes (comportant plusieurs verbes). Elle correspond à l'idée que l'on n'essaye pas de trouver une analyse complète cohérente de la phrase mais à retrouver les actants potentiels de chacun de ses verbes, indépendamment. Certaines analyses peuvent être, par conséquent, incompatibles (un même groupe syntaxique peut être actant de deux verbes distincts de la phrase). Ce choix est motivé par des considérations d'efficacité. En effet, étant donné le relâchement syntaxique mis en œuvre, la recherche d'analyses complètes aboutit à une multiplication déraisonnable de solutions, et par conséquent une augmentation déraisonnable de la taille des automates produits.

2.2 Exemple

Nous avons représenté ci-dessous le résultat du traitement du graphe de reconnaissance correspondant à l'énoncé *oui donc j' aimerais réserver maintenant du onze au euh quatorze mai c'est-à-dire trois nuits la même prestation à carcassonne*. Le graphe est traité successivement par le module lexical, l'étiqueteur morpho-syntaxique, l'analyseur morphologique, l'analyseur syntaxique partiel puis, finalement par l'analyseur en dépendances partiel. Ce sont les dix solutions de meilleur poids que l'on a reportées dans le tableau ci-dessous[4].

[2]Pour des raisons de lisibilité, nous avons omis un certain nombre de détails techniques concernant la structure des automates T_{F_i} et T_S. En particulier, avant l'opération de composition, les transitions de T_S étiquetées par un groupe verbal dont la tête est associée à n cadres valenciels est remplacée par n transitions étiquetées par l'identifiant du cadre valenciel. D'autres détails ne présentant pas d'interêt fondamental sont passés sous silence.

[3]Dans la pratique, on effectue des n meilleurs chemins de l'automate produit lors de l'opération de composition. La formule est donc : $\cup_{i=1}^{n} \mathrm{nbest}(T_S \circ T_{F_i}, n)$ où $\mathrm{nbest}(A, n)$ est l'opérateur produisant l'automate composé des n chemins de meilleur poids de l'automate A.

[4]La présence incongrue du mot *stade* dans *quatorze mai stade* provient d'une erreur de reconnaissance : substitution de *c'est à dire* par *stade*.

n	Cadre	Sujet	Verbe	Objet direct	Complément locatif
1	n0V	*j'*	*aimerais réserver*	∅	∅
2	n0Vn1	*j'*	*aimerais réserver*	*quatorze mai stade*	∅
3	n0Vn1	*j'*	*aimerais réserver*	*trois nuits*	∅
4	n0Vn1locn2	*j'*	*aimerais réserver*	*quatorze mai stade*	*à carcassonne*
5	n0Vn1locn2	*j'*	*aimerais réserver*	*la même prestations*	*à carcassonne*
6	n0Vn1locn2	*j'*	*aimerais réserver*	*trois nuits*	*à carcassonne*
7	n0Vn1	*j'*	*aimerais réserver*	*la même prestations*	∅
8	n0Vlocn1	*j'*	*aimerais réserver*	∅	*à carcassonne*
9	n0Vn1	*j'*	*aimerais réserver*	*stade*	∅
10	n0Vn1locn2	*j'*	*aimerais réserver*	*la même*	*à carcassonne*

Les dix solutions représentées dans le tableau correspondent à dix instanciations de cadres valenciels associés au verbe *réserver*. Quatre cadres valenciels ont été sélectionnés, n0V (intransitif), n0Vn1 (transitif direct), n0Vn1locn2 (objet direct et complément locatif prépositionnel) et n0Vlocn1 (complément locatif). Le classement de ces solutions est réalisé d'après la combinaison des poids donnés par les différents modules successifs. Il s'agit d'une fonction de coût complexe, combinaison linéaire des fonctions de coût de chaque module.

Nous aurions aimé voir en premières positions de la liste les solutions 5 et 6 qui peuvent être considérées comme des analyses correctes de l'énoncé. Leur positionnement en 5ème et 6ème rang illustre l'imperfection de la fonction de coût. Mais il montre aussi qu'il est illusoire d'espérer définir une fonction de coût générique optimale. En particulier, notre fonction de coût n'implémente aucune contrainte de sélection (ou préférence lexicale) qui favoriserait, par exemple, le groupe nominal *trois nuits* par rapport à *quatorze mai stade* comme objet direct de *réserver*. De telles préférences sont étroitement liées à un domaine sémantique et à un cadre applicatif particuliers, c'est pourquoi nous considérons qu'elles sont du ressort d'un module spécifique à une application. Ce dernier est décrit dans la section suivante, il se présente sous la forme d'un réordonnanceur (*reranker*) qui prend en entrée n solutions de l'analyseur et les réordonne en fonction de connaissances propres à l'application.

3 Réordonnancement des solutions de l'analyseur

Comme nous l'avons décrit dans la section précédente, les traitements génériques ont permis de construire, pour les sorties d'un SRAP correspondant à un énoncé donné, plusieurs instanciations de cadres de valence associés aux verbes de l'énoncé considérés pertinents pour la tâche visée (*verbes cibles* dans la suite de l'article). Les cadres de valences ainsi produits constituent l'entrée d'un processus de réordonnancement qui intègre des contraintes propres à une application particulière.

Ce processus est basé sur un classifieur qui est entraîné pour sélectionner les analyses valides parmi l'ensemble des analyses produites. Nous utilisons un classifieur à large marge spécialisé dans le traitement de données textuelles, ICSIBOOST[5], basé sur un algorithme de *boosting* (Schapire & Singer, 2000) de classifieurs simples (des arbres de décision à 1 niveau de profondeur sur la présence ou l'absence de n-grammes de mots). L'entraînement de ce classifieur est effectué de la manière suivante :

1 - Tout d'abord un corpus d'apprentissage composé de segments de parole transcrits manuel-

[5]http ://code.google.com/p/icsiboost/

lement est traité par l'analyseur.

2 - Pour chaque segment, et pour chaque verbe cible, la liste de n meilleurs cadres de valences produits est présentée à un juge humain qui valide la ou les analyses correctes, lorsqu'elles existent. A l'issue de cette étape de validation manuelle, chaque cadre valenciel est étiqueté comme correct ou erroné.

3 - Tous les cadres valenciels produits sur le corpus d'apprentissage sont alors regroupés et chacun d'eux est représenté par la séquence de mots du segment enrichis de la fonction syntaxique du constituant auquel il appartient ou *NULL* si le mot n'appartient à aucun constituant. Le classifieur ICSIBOOST est alors entraîné pour séparer les exemples jugés corrects des exemples jugés incorrects.

Lors du traitement d'un segment de parole, le classifieur réévalue chaque cadre de valence instancié produit par l'analyseur robuste et produit un nouveau classement de la liste de n-meilleures hypothèses. A l'issue de cette phase de réordonnancement, l'hypothèse ou les hypothèses ayant les meilleurs scores au sens du classifieur sont choisies comme solution.

Dans l'exemple de la section précédente, le juge aura marqué les deux hypothèses de rang 5 et 6 comme correctes. Il faut noter que cette étape d'adaptation manuelle à une tâche est bien plus légère que l'annotation manuelle de corpus ou l'adaptation de grammaire. Elle est cependant limitée par la nature de la tâche de réordonnancement : elle ne peut proposer une solution qui n'a pas été créée par l'analyseur.

4 Cadre expérimental et évaluation

Le cadre applicatif choisi est issu du corpus de dialogue MEDIA (Bonneau-Maynard *et al.*, 2005) développé lors du projet Technolangue homonyme. Ce corpus a été enregistré selon un protocole de type *Magicien d'Oz* simulant un serveur vocal téléphonique permettant la réservations d'hôtels. Huit catégories de scénario ont été définies correspondant à des degrés de complexités différents. Le corpus compte 1250 dialogues enregistrés auprès de 250 interlocuteurs et représente environ 70 heures de parole.

L'analyseur de la section 2 a été appliqué sur deux versions des transcriptions du corpus MEDIA : les transcriptions manuelles de référence et les transcriptions automatiques obtenues à partir des fichiers audio du corpus grâce au système de reconnaissance automatique de la parole SPEERAL (Nocera *et al.*, 2002). Dans cette première étude nous avons volontairement restreint le champs expérimental en ne considérant qu'un seul verbe *cible*, le verbe : "*réserver*", évidemment central dans une tâche de réservation hotelière.

Le corpus utilisé dans cette étude a été découpé en trois parties : un corpus d'apprentissage constitué des lots 1,2,3 et 4 du corpus MEDIA ; un corpus de développement constitué du "*test à blanc*" de la campagne MEDIA et un corpus de test correspondant au corpus "*test Hors-Contexte*" de la campagne MEDIA. Les caractéristiques des trois parties sont décrites ci-dessous :

Corpus	**Apprentissage**	**Développement**	**Test**
Nb de dialogues	727	79	208
Nb de tours de parole (utilisateur)	12988	1265	3524
Nb d'occurrences du verbe *réserver*	659	73	187
Taux d'erreur/mot	14.5	25.3	27.4

Les trois corpus ont été traités par l'analyseur et à chaque tour de parole a été associée la liste d'hypothèses de dépendance (éventuellement vide) pour le verbe cible. A l'issue de cette analyse, les listes d'hypothèses ont été manuellement vérifiées pour marquer les cadres de valences corrects. Le classifieur chargé du réordonnancement des hypothèses a été entraîné sur ce corpus d'apprentissage étiqueté. Les paramètres de l'algorithme (nombre d'itérations, taille des n-grammes pour les classifieurs simples, ...) ont été ajustés sur le corpus de développement. Rappelons que seul le réordonnanceur a été entraîné sur le corpus d'apprentissage. L'analyseur syntaxique, lui, n'a subi aucune adaptation. Enfin l'évaluation a été faite sur le corpus de test.

Les performances sont exprimées par les mesures de rappel, précision et F-mesure en utilisant deux mesures d'évaluation. Dans la première, une hypothèse est jugée correcte si le cadre de valence sélectionné pour le verbe est correct ainsi que tous les actants (leur fonction syntaxique et la séquence de formes qui les constituent) La seconde mesure (représentée entre parenthèses dans le tableau) est moins sévère, elle n'évalue que la détection des actants d'un verbe (leur fonction syntaxique et la séquence de formes qui les constituent), en faisant abstraction du cadre de valence.

Trois systèmes ont été évalués :

– le système *Oracle* consiste à choisir l'hypothèse correcte, si elle existe, dans la liste des n-meilleures hypothèses produites par l'analyseur. Il correspond à la borne maximale que l'on peut atteindre à l'issue du réordonnancement.
– le système *Baseline* se contente de choisir la première hypothèse produite par l'analyseur.
– le système *Réordonnancement* réordonne les hypothèses de l'analyseur avec le classifieur présenté dans la section précédente et entraîné sur le corpus d'apprentissage MEDIA.

Les résultats sont donnés ci-dessous :

Corpus	**trans. manuelle**						**trans. auto**					
système	précision		rappel		F-mesure		précision		rappel		F-mesure	
Oracle	90.0		90.0		90.0		63.2		63.2		63.2	
Baseline	42.8	(88.8)	38.3	(76.6)	40.4	(82.4)	34.7	(78.2)	29.2	(63.5)	31.6	(70.1)
Reclas.	59.9	(91.1)	70.8	(90.2)	64.9	(90.6)	45.9	(80.4)	53.1	(76.0)	49.2	(78.2)

Les résultats présentés ci-dessus permettent de tirer un certain nombre de conclusions. La couverture de l'analyseur sur des entrées transcrites manuellement sont bonnes (90% de F-mesure). L'analyseur se comporte donc bien sur des retranscriptions manuelles de l'oral. Cette couverture chute à 63% pour des transcriptions automatiques, illustrant la sensibilité de l'analyseur aux erreurs du SRAP. La médiocrité des résultats du système baseline peut être interprétée de deux manière : la fonction de coût de l'analyseur est globalement mauvaise (certains choix sont syntaxiquement aberrants) ou bien elle n'est pas adaptée au corpus traité (certains choix sont syntaxiquement raisonnables, mais sémantiquement incorrects). Une inspection rapide des résultats montre que la réalité se situe entre les deux : la fonction de coût de l'analyseur peut être améliorée mais certains cas relèvent de la sémantique. Enfin, la comparaison de la baseline et du système de réordonnancement montre le rôle important du réordonnanceur (amélioration de 55% de la F-mesure sur les transcriptions automatiques), mais qu'il reste une marge de progression importante (28.4%). Ces tendances sont aussi valables pour la seconde mesure d'évaluation, avec des performances globales plus satisfaisantes (78.2%) pour une tâche qui reste intéressante dans le cadre de la compréhension de parole.

5 Conclusions

Nous avons proposé dans cet article un modèle de traitement syntaxique de l'oral spontané qui se décompose en deux étapes. Une analyse syntaxique axée sur la détection de cadre valenciels, suivie d'une étape de réordonnancement des résultats de l'analyse. La première est elle même composée d'une analyse syntagmatique partielle, suivie d'une étape de détection de cadres valenciels. Ces deux dernières ont été adaptées au bruit inhérent aux graphes produits par des SRAP. Mais cette adaptation doit être améliorée. En particulier la détection de cadres valenciels repose sur un relâchement extrême des contraintes syntaxiques, qui pourrait être modéré. D'autre part, l'analyse syntagmatique manque de robustesse face aux phénomènes propres à l'oral. Deux voies sont envisagées : la modification de la grammaire syntagmatique et le prétraitement des graphes produits par le SRAP afin d'en éliminer certains phénomènes (hésitations, certaines répétitions).

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In A. ABEILLÉ, Ed., *Treebanks*. Dordrecht : Kluwer.

ABNEY S. (1996). Partial parsing via finite-state cascades. In Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information, Prague, Czech Republic.

ANTOINE J.-Y., GOULIAN J. & VILLANEAU J. Quand le TAL robuste s'attaque au langage parlé : analyse incrementale pour la compréhension de la parole spontanée. In *TALN*.

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The Berkeley FrameNet project. In *COLING/ACL-98*, p. 86–90.

BONNEAU-MAYNARD H., ROSSET S., AYACHE C., KUHN A. & MOSTEFA D. (2005). Semantic annotation of the french media dialog corpus. In *Eurospeech*, Lisboa, Portugal.

CRABBÉ B. (2005a). Grammatical development with xmg. In *Logical Aspects of Computational Linguistics*.

CRABBÉ B. (2005b). *Représentation informatique de grammaires fortement lexicalisées, application à la grammaire d'arbres adjoints*. PhD thesis, Université Nancy 2.

LEHMANN S. *et al*. (1996). TSNLP : Test suites for natural language processing. In *Proceedings of the 16th conference on Computational linguistics*, p. 711–716.

NOCERA P., LINARES G. & MASSONIE D. (2002). Principes et performances du décodeur parole continue Speeral. In *Proc. Journées d'Etude sur la Parole (JEP)*.

SAGOT B., CLÉMENT L., ÉRICE VILLEMONTE DE LA CLERGERIE & BOULLIER P. (2006). The lefff 2 syntactic lexicon for french : architecture, acquisition, use. In *LREC*.

SCHAPIRE R. E. & SINGER Y. (2000). BoosTexter : A boosting-based system for text categorization. *Machine Learning*, **39**, 135–168.

SENEFF S. (1992). TINA : A natural language system for spoken language applications. *Computational Linguistics*, **18**(1), 61–86.

VAN DEN EYNDE K. & BLANCHE-BENVENISTE C. (1978). Syntaxe et mécanismes descriptifs : présentation de l'approche pronominale. *Cahiers de lexicologie*, **32**, 63–104.

VAN DEN EYNDE K. & MERTENS P. (2003). La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, **13**, 63–104.