L'analyseur syntaxique Cordial dans Passage[1]

Dominique Laurent, Sophie Nègre, Patrick Séguéla (1)

(1) Synapse Développement
dlaurent, sophie.negre, patrick.seguela@synapse-fr.com

Résumé : Cordial est un analyseur syntaxique et sémantique développé par la société Synapse Développement. Largement utilisé par les laboratoires de TALN depuis plus de dix ans, cet analyseur participe à la campagne Passage ("Produire des Annotations Syntaxiques à Grande Échelle"). Comment fonctionne cet analyseur ? Quels résultats a-t-il obtenu lors de la première phase d'évaluation de cette campagne ? Au-delà de ces questions, cet article montre en quoi les contraintes industrielles façonnent les outils d'analyse automatique du langage naturel.

Abstract : Cordial is a syntactic and semantic parser developed by Synapse Développement. Widely used by the laboratories of NLP for over ten years, this analyzer is involved in the Passage campaign ("Producing Syntactic Annotations on a Large Scale"). How does this parser work? What were the results obtained during the first phase of the evaluation? Beyond these issues, this article shows how the industrial constraints condition the tools for Natural Language Procesing.

Mots-clés : Analyse syntaxique, analyse sémantique, évaluation, Passage
Keywords: Parsing, semantic analysis, evaluation

[1] cf. http://atoll.inria.fr/passage/eval1.fr.html, et http://atoll.inria.fr/passage/articles.fr.html

1 Introduction

1.1 Cordial

Cordial (acronyme de CORrecteur D'Imprécisions et Analyseur Lexico-sémantique) est un analyseur syntaxique conçu à l'origine pour la correction orthographique et grammaticale. Développé au début des années 90 mais constamment maintenu et enrichi depuis, Cordial est le fondement de nombreux développements : composants linguistiques de nettoyage automatique de texte, extracteur de mots-clés et de phrases-clés, extracteur de thèmes et de concepts, extracteur de terminologie et d'entités nommées, moteur de question-réponse.

Développé sous Windows, Cordial est également disponible sous Linux, en librairie d'analyse et de correction. Une version spécifique pour la recherche, "Cordial Analyseur", est utilisée par plus d'une centaine de laboratoires dans le monde. Conçu pour la langue française, Cordial devient bilingue, l'analyseur syntaxique et de nombreuses ressources ayant été progressivement adaptées pour la langue anglaise.

1.2 Easy et Passage

De 1994 à 1998 a eu lieu la campagne d'évaluation GRACE (Grammaire et Ressources pour les Analyseurs de Corpus et leur Évaluation). Notre analyseur, tout juste commercialisé, a obtenu d'excellentes notes lors de la phase préparatoire. Il n'a malheureusement pas pu participer à l'évaluation finale, celle-ci ayant été prévue pour des systèmes sous Unix et notre système étant à l'époque uniquement sous Windows.

Dans le cadre du projet Technolangue-Evalda, une campagne d'évaluation EASY (pour Évaluation des Analyseurs SYntaxiques) s'est déroulée entre 2003 et 2006. Plus ambitieuse que GRACE, cette campagne visait à évaluer l'analyse syntaxique en composants et en relations. Cordial a également participé à cette campagne, avec d'excellents résultats fournis plus bas (§ 3).

De 2007 à 2009, dans le cadre du projet PASSAGE ("Produire des Annotations Syntaxiques À Grande Échelle"), une pré-évaluation a eu lieu fin 2007 et une évaluation aura lieu cette année. Nous participons également à cette campagne.

2 L'analyseur syntaxique Cordial

Comme mentionné supra, Cordial est un analyseur conçu pour la correction. Afin d'analyser le mieux possible un texte comportant éventuellement des fautes d'accord ou d'autres fautes, il est difficilement envisageable d'utiliser une grammaire formelle, qui fait le plus souvent appel à des règles d'appariement de groupes privilégiant le genre et le nombre. Nous avons donc associé à des règles générales et très peu liées aux attributs de genre et de nombre (analyseur à relâchement de contraintes) un ensemble d'outils statistiques, en particulier pour effectuer la désambiguïsation grammaticale.

2.1 La désambiguïsation grammaticale

Dans un dictionnaire, la proportion de mots polysyntaxiques, par exemple noms et adjectifs (*bonne, politique, vert...*), formes verbales et noms (*avions, boucher, foule, été...*), est faible. Elle représente moins de 5% des entrées, donc des lemmes, même pour un dictionnaire classique, et moins de 2% des formes (si on ne considère pas les ambiguïtés entre formes verbales). Dans la réalité de la langue, cette proportion est considérable puisqu'en moyenne plus de 50% des mots d'une phrase sont ambigus grammaticalement. Cette proportion très élevée résulte d'une part de l'ambiguïté de nombreux mots-outils (*le, la, les*, déterminants et pronoms personnels, *des, du*, déterminants simples et contractés, etc.) et, d'autre part, du fait que les mots les plus courants sont souvent les plus ambigus.

Cette ambiguïté constitue un problème majeur pour tout analyseur syntaxique car toute erreur de désambiguïsation grammaticale est très pénalisante pour les étapes ultérieures : groupage, recherche des relations, analyse sémantique. Ainsi dans l'exemple classique "*le pilote ferme la porte*", en admettant que "*le pilote*" est considéré comme déterminant + nom et non comme personnel + verbe le choix de l'adjectif pour "*ferme*" induit personnel + verbe pour "*la porte*" alors que le choix de la forme verbale induit déterminant + nom pour le groupe "*la porte*".

La désambiguïsation grammaticale se situe dans notre analyseur après la consultation des dictionnaires et la correction orthographique. La consultation des dictionnaires affecte à chaque mot l'ensemble des types grammaticaux possibles, y compris les différentes formes verbales. La correction orthographique, remplace les mots inconnus par la meilleure suggestion. Ce remplacement n'est effectué que si la première suggestion a une probabilité nettement supérieure à la seconde et selon le paramétrage utilisateur. Ce traitement est optionnel. Les entités nommées figurant dans nos dictionnaires sont également groupées à cette étape. La détection des entités nommées s'appuie sur les dictionnaires très complets développés pour Cordial, à notre connaissance, les plus complets en français (plus de un million de lemmes pour plus de 2 millions de mots, noms communs et noms propres réunis, plus de 100 000 expressions). De plus, avant la désambiguïsation, un premier groupement des expressions est effectué. À ce stade, les expressions prépositionnelles sont privilégiées car leur non-détermination pèse fortement sur les analyses ultérieures. Pour la totalité des verbes et pour un grand nombre d'autres mots (au total environ 150 000 lemmes), un dictionnaire "grammatical" fournit de 8 (pour un adverbe) à 147 traits (pour un verbe courant), traits sémantiques (type sémantique du sujet, du COD, du COI, par exemple) ou traits syntaxiques (complétive possible, infinitive peut suivre, etc.)

L'application de quelques dizaines de règles simples permet d'éliminer environ 12 % des ambiguïtés grammaticales. Ainsi "*le*" ou "*la*" précédé d'un pronom personnel sujet ("*je*", "*il*"...) non relié à une forme verbale le précédant est toujours un pronom personnel, "*mien*" ou "*sien*" précédé d'un déterminant est toujours un pronom possessif, etc. Ces règles sont toutefois moins simples que décrit ci-dessus car la robustesse nécessaire d'un correcteur impose ses lois à l'analyseur. Par exemple le trait d'union reliant le pronom personnel sujet à la forme verbale le précédant a pu être oublié, il faut donc vérifier si ce pronom personnel est précédé d'une forme verbale éventuelle, etc.

Après l'application de ces règles, la proportion de mots ambigus grammaticalement reste encore très importante. Pour résoudre ces ambiguïtés, nous utilisons une table de bigrammes et une table de trigrammes. La table de bigrammes croise 70 catégories grammaticales correspondant aux différents types grammaticaux avec leur genre et leur nombre s'ils existent. Pour les formes conjuguées, seule la personne est prise en compte (le mode et le temps sont

désambiguïsés ultérieurement). 6 catégories concernent la ponctuation (point ou fin de phrase, virgule, point virgule, deux points, parenthèse ouvrante ou fermante). Pour certaines catégories peu fréquentes, comme le pronom possessif ou l'adjectif indéfini, aucune différenciation n'est effectuée, ni de genre, ni de nombre. La table de trigrammes croise 10 catégories soit 10x10x10 (en fait 10x10x11 avec les effectifs cumulés). Ces 10 catégories sont les grandes bases de désambiguïsation : adjectifs, adverbes, déterminants, conjonctions, prépositions, noms, pronoms, verbes, ponctuations faibles, ponctuations fortes.

Ces tables ont été construites en 1992 à partir de petits corpus annotés manuellement d'environ 50 000 mots. Depuis ces corpus ont été enrichis, à travers les campagnes d'évaluation mais également par correction manuelle de nos sorties et portent maintenant sur plus de deux millions de mots. Les tables de bigrammes et trigrammes sont mises à jour au fur et à mesure de l'évolution de ce corpus annoté (en fait tous les deux ans en moyenne). Les textes utilisés sont des textes "propres", a priori exempts de fautes, ce qui pèse parfois sur les choix effectués lors de l'analyse de textes dégradés. Ainsi de l'oubli de l'accent grave sur "*a*" qui, en passant ce mot d'une catégorie de préposition à une forme verbale ou un nom, contamine la désambiguïsation des termes voisins.

La table de bigrammes est appliquée sur le mot précédent et le mot ambigu, ainsi que sur le mot ambigu et le mot suivant. La table de trigrammes est appliquée sur les deux mots précédents et le mot ambigu, sur le mot précédent, le mot suivant et le mot ambigu, sur le mot ambigu et les deux mots suivants. Les probabilités obtenues sont combinées selon une formule prenant également en compte les ambiguïtés possibles des mots précédents et suivants ainsi que la fréquence dans la langue des différentes formes de chaque mot, par exemple le fait que la forme verbale "*foule*" est beaucoup moins fréquente que le substantif ou le fait que la forme verbale "*été*" est nettement plus fréquente que le substantif. Ces statistiques de fréquence ont été élaborées au départ à partir d'un corpus de 500 millions de mots, elles ont été revues l'an dernier à partir d'un corpus de 1,2 milliard de mots. La désambiguïsation étant effectuée de gauche à droite, un poids plus élevé est donné aux probabilités avec le ou les mots précédents.

Tous les mots sont alors "désambiguïsés" grammaticalement. En fait, le tableau des formes grammaticales possibles pour chaque mot est conservé jusqu'à la fin de l'analyse, de même que la probabilité de chaque forme grammaticale, permettant des changements jusqu'à la phase de détection des relations. Un module spécifique traite alors les ambiguïtés "lourdes", c'est-à-dire celles constituées d'au moins deux mots ambigus, pour lesquelles les tables sont parfois insuffisantes et qui nécessitent la prise en compte d'un contexte supérieur à quatre mots. C'est le cas des couples adjectif/nom, nom/adjectif ("*bonne alerte*"), déterminant/nom ou personnel/verbe ("*l'aide*", "*les avions*").

Une première phase de désambiguïsation sémantique est alors effectuée. Elle se base essentiellement sur le contexte gauche et droit pour affecter une probabilité à chacun des sens possibles des mots polysémiques. Nous prenons en compte environ 25 000 sens pour 9 000 mots polysémiques, ce qui est un peu inférieur à un dictionnaire papier, le but étant de séparer des sens correspondant à des usages syntaxiques hétérogènes et surtout à des concepts nettement différenciés. Ainsi "abdomen" est monosémique pour Cordial alors que de nombreux dictionnaires distinguent la région inférieure du corps des mammifères et la partie postérieure du corps des arthropodes.

2.2 Le groupage en constituants

La première phase du groupage en constituants consiste à regrouper les expressions (verbales, nominales, adjectivales, adverbiales...). Ce regroupement utilise la probabilité de cohérence de l'expression (codée de 1 à 9 dans nos lexiques) et réapplique les tables de bigrammes et de trigrammes entre l'expression potentielle et les mots ou expressions précédents et suivants. De fait 92 % des expressions potentielles sont regroupées mais certaines expressions à faible cohérence comme "*bien que*" sont moins fréquemment groupées.

Le groupage en constituants s'effectue en quatre phases : constitution des groupes verbaux, puis des groupes nominaux, puis des groupes adjectivaux, enfin traitement des mots résiduels, en particulier des adverbes, raccordés ou non à un groupe existant. Une fois les groupes de base constitués, un module recherche les dépendances de groupes. Il raccorde par exemple "*des fêtes*" à "*comité*" et "*du comité*" à "*président*" dans "*président du comité des fêtes*". Rappelons que le groupage en constituants s'effectue avec une prise en compte minimale du genre et du nombre, le but étant de corriger d'éventuelles erreurs d'accord.

2.3 La détermination des relations

Les groupes étant délimités et définis, il devient possible de déterminer les relations entre ces groupes. Avant cela, il est toutefois nécessaire de découper la phrase en propositions. Cette phase est beaucoup plus complexe qu'il n'y paraît, les délimiteurs de proposition comme "*que*" ou "*et*" pouvant également être souvent des unificateurs de groupes, et les propositions pouvant être constituées de différents fragments ("*l'homme qui vous parle et la femme qui vous regarde sont de fait mari et femme*"). Cette découpe en propositions s'effectue à partir des groupes verbaux à la fois vers la gauche et la droite, elle prend en compte les contraintes syntaxiques et sémantiques des verbes, par exemple le type sémantique du sujet, de l'attribut, du COD. Le typage des propositions (principale, indépendante, incise, etc.) est effectué durant cette phase de découpe, les délimiteurs potentiels étant conservés pour d'éventuelles corrections ultérieures.

Après la relation sujet-verbe, c'est la relation verbe-attribut qui est recherchée, si elle existe. Ces deux relations sont en effet les deux plus importantes sources de fautes de grammaire, hors accord interne des groupes en genre et en nombre. Puis, les autres relations, COD, COI, apposition, etc. sont recherchées pour chacun des groupes. Avant cette recherche, une détermination de la nature réelle des déterminants "*de la*" "*du*" et "*des*" est effectuée car elle conditionne la catégorisation en objet direct ou indirect ("*il vend de la toile de Jouy*", "*il vend de bourg en bourg*"). Un traitement final vérifie l'affectation des adjectifs entre ponctuations ou en fin de groupe nominal composé afin de déterminer au mieux le rattachement de ces adjectifs. Enfin est alors effectuée la désambiguïsation des modes et temps verbaux (indicatif ou subjonctif, par exemple), même si certaines formes comme l'impératif sont souvent déjà désambiguïsées à ce stade (absence de sujet).

La détermination des relations s'appuie avant tout sur la nature des groupes constituant la proposition et leur position dans la phrase, mais elle prend en compte l'ensemble des données stockées dans notre dictionnaire grammatical. Enfin une deuxième phase de la désambiguïsation sémantique est alors effectuée, qui prend en compte l'ensemble des contraintes syntaxiques et sémantiques des étapes antérieures. Cette seconde désambiguïsation est encore susceptible d'amener une révision des mises en relation et parfois même de la désambiguïsation grammaticale.

2.4 En fin d'analyse

Lorsque l'analyseur est utilisé pour la correction, application première de Cordial, c'est à ce stade que se situe l'application des règles d'accord intra-groupes ou inter-groupes. Certaines règles font l'objet de modules spécifiques, en particulier l'accord du participe passé et la concordance des temps.

La résolution des anaphores et la détection des entités nommées dépassent un peu le cadre de l'analyseur proprement dit, mais se situent après cette phase de correction. Les référents potentiels sont collationnés au fil de l'analyse et la résolution des anaphores pronominales et adjectivales (adjectifs possessifs) utilise ces référents pour déterminer auquel se réfère l'anaphore, en fonction de critères essentiellement sémantiques et statistiques, sachant par exemple que les pronoms démonstratifs réfèrent le plus souvent à des référents très proches, que le pronom personnel "en" peut référer à une phrase entière plutôt qu'à un groupe nominal ou une entité nommée, etc.

La détection des entités nommées ne figurant pas dans nos dictionnaires s'appuie sur l'analyse syntaxique et les données sémantiques réunies sur chaque terme. Elle permet de regrouper de nombreuses entités ne figurant pas dans nos dictionnaires et opère surtout une désambiguïsation de ces entités nommées ("France" réfère-t-il à une femme, à un pays comme lieu ou comme organisation, à un bateau, etc.).

Les utilisations industrielles de notre analyseur imposent des contraintes fortes de robustesse et de rapidité. Conçu pour la correction, Cordial est particulièrement adapté pour l'analyse de corpus dégradés et de corpus oraux. L'impératif de rapidité a également imposé ses lois sur la conception d'ensemble de l'analyseur. La récursivité et les réanalyses sont pratiquement bannies, même si les nombreuses variables probabilistes permettent des rectifications en cours d'analyse. Toutes les routines ont été écrites en langage C optimisé. La vitesse d'analyse est d'environ 10 000 mots par seconde, en intégrant les fonctions optionnelles de correction, de recherche des anaphores et de marquage des entités nommées.

3 Easy

Participer à une évaluation d'analyseurs syntaxiques, c'est évaluer la valeur de son analyseur mais aussi évaluer l'aptitude de ses développeurs à se conformer à une norme. Pour Easy et pour Passage, cette norme est sophistiquée et complexe, assez éloignée du format de travail de Cordial. Ainsi l'évaluation définissait un petit lexique d'expressions acceptées, très éloigné en taille de notre dictionnaire, ce qui nous a obligé à effectuer des traitements spécifiques conséquents, des modules entiers de notre analyseur devant être neutralisés ou modifiés.

Nos résultats sont excellents pour l'analyse en constituants : 1er en précision, 1er en rappel et 1er en f-mesure avec un taux moyen de 0,89 (système A10 dans les tableaux fournis par Paroubek 2008). Pour l'analyse en relations, nous obtenons la meilleure f-mesure (0,58), un système ayant une meilleure précision et un autre système ayant un meilleur rappel. Notre système traitait l'ensemble des corpus et des relations. Il s'est montré très régulier pour les corpus, ne perdant qu'environ 5% de précision et rappel en constituants, et moins de 10% de précision et rappel en relations pour les corpus de mails et oral, les plus difficiles.

Notons ici que la complexité du codage en constituants et en relations s'est traduite par des taux d'erreur élevés tant dans les corpus de développement que dans les corpus de tests, taux

évalués à près de 5% sur les constituants et au moins 20 % sur les relations. Pour la campagne Passage, ces corpus ont été révisés et le guide d'annotation a été révisé mais les taux d'erreurs restent non négligeables.

4 Pré-campagne Passage

Cette campagne vient à la suite des campagnes GRACE et EASY et reprend plusieurs des protocoles d'évaluation de cette dernière campagne, avec des corpus sensiblement différents et plus étendus.

4.1 Objectifs et déroulement

Les principaux objectifs de cette campagne soutenue par l'ANR sont les suivants :

- évaluer les analyseurs français ;
- améliorer l'exactitude et la robustesse des analyseurs français sur des corpus à grande échelle (270 millions de mots) ;
- exploiter les annotations syntaxiques résultantes pour créer une ressource linguistique plus riche et plus étendue : un treebank pour le Français.

Six types de constituants ont été choisis pour cette campagne : Groupe nominal, Noyau verbal, Groupe adjectival, Groupe adverbial, Groupe prépositionnel et Groupe prépositionnel à noyau verbal.

Les relations à relever sont les suivantes : dépendance sujet-verbe (SUJ-V), dépendance auxiliaire-verbe (AUX-V), objet direct (COD-V), autres compléments du verbe (CPL-V), modifieurs du verbe (MOD-V), subordonnés (COMP), attribut du sujet ou de l'objet (ATB-SO)**,** modifieur du nom (MOD-N)**,** modifieur de l'adjectif (MOD-A)**,** modifieur de l'adverbe (MOD-R)**,** modifieur de la préposition (MOD-P)**,** coordination (COORD)**,** apposition (APPOS) et juxtaposition (JUXT).

Lors de la première évaluation de la campagne Passage qui s'est déroulée fin 2007, 10 analyseurs ont fourni leurs résultats en constituants et seulement 7 ont pu être évalués sur les relations.

4.2 Cordial dans la campagne Passage1

Pour cette campagne, un serveur d'évaluation a été mis en place, qui permet aux participants de tester leur système sur un corpus d'entraînement. Ce serveur d'évaluation, en permettant autant d'itérations que désiré et durant plusieurs mois, a été et reste d'un grand intérêt pour l'amélioration des systèmes. A titre indicatif, après un mois et demi d'améliorations sur la délimitation des constituants (novembre-décembre 2007), l'analyseur Cordial offre un taux de performance compris entre 96% et 99% selon les corpus. Pour les relations, après un mois d'évaluations comparatives (mi-janvier à mi-février 2008), l'analyseur Cordial offre des performances comprises entre 70% et 80% selon les corpus. Ces performances sont naturellement très différentes selon le type des relations.

L'analyseur Cordial a de bonnes ou de très bonnes performances (80% à 100%) sur les relations liées aux accords (sujet-verbe, verbe-attribut, auxiliaire-verbe, verbe-cod, modifieur de nom), il obtient par contre des résultats inférieurs sur d'autres types de relations comme les

appositions ou les juxtapositions. Une autre conséquence de l'amélioration grâce au serveur d'évaluation est la convergence progressive des courbes de précision et de rappel : comme l'ensemble des systèmes, l'analyseur Cordial offrait au départ une précision supérieure au rappel. Cette différence a été progressivement gommée.

Une évaluation a eu lieu fin 2007. Lors de cette évaluation, Cordial a obtenu des résultats similaires à ceux de la campagne EASY : 1er en précision, en rappel et en f-mesure pour les composants, 1er en rappel, 2e en précision et en f-mesure :

Constituants			**Relations**		
Précision	Rappel	**F-mesure**	Précision	Rappel	**F-mesure**
0.96	0.97	**0.96**	0.69	0.65	**0.67**

Figure 1 : résultats de la pré-campagne Passage (arrondis à 2 décimales)

On notera que nos résultats sont nettement meilleurs en constituants que lors d'EASY, le taux d'erreurs étant inférieur à 4% contre 11% lors d'EASY. Concernant les relations, nos résultats sont également meilleurs que pour EASY : plus de 10 % d'amélioration.

	Constituants			**Relations**		
	Précision	Rappel	**F-mesure**	Précision	Rappel	**F-mesure**
Corpus complet	0,96	0,97	0,96	0,69	0,65	0,67
Elda	0,97	0,98	0,97	0,76	0,69	0,72
Le Monde	0,96	0,97	0,96	0,66	0,63	0,64
Littéraire	0,96	0,96	0,96	0,70	0,65	0,68
Mail	0,91	0,95	0,93	0,65	0,57	0,60
Médical	0,95	0,96	0,95	0,67	0,67	0,67
Delic	1,00	1,00	1,00	1,00	0,50	0,67
Parlement	0,96	0,97	0,97	0,68	0,67	0,68
Questions	0,97	0,98	0,98	0,67	0,66	0,67

Figure 2 : résultats par corpus de la pré-campagne Passage (arrondis à 2 décimales)

On remarquera aussi une certaine irrégularité des résultats de Cordial selon le type de corpus à analyser. Ainsi, en ce qui concerne l'analyse des constituants, les résultats sont assez largement inférieurs à la moyenne pour les corpus Oral, Mail et Médical. On peut expliquer cet état de fait en considérant que les corpus Oral et Mail présentent de nombreuses constructions peu grammaticales, parfois une absence ou une profusion de ponctuations, des disfluences, etc . Les relatives difficultés d'analyse concernant le langage de spécialité (Médical), tendent à prouver qu'il possède des tournures propres. À noter que, pour les constituants, les résultats fournis par Cordial, qui annote plus de 10 000 mots par seconde, sont très voisins du taux de concordance entre des annotateurs humains, annotant quelques mots par minute.

Une analyse des erreurs commises par type de constituants montre que les principales erreurs sont liées à la confusion entre adjectif et participe passé ainsi qu'à de mauvais raccordements de groupes nominaux soit vers d'autres groupes nominaux soit vers des verbes.

En ce qui concerne les relations, les corpus les plus difficiles à analyser par Cordial sont l'Oral, les Mails et la Littéraire, ce qui n'est guère surprenant. En effet, la difficulté d'identifier de manière précise les constituants dans les phrases pour les corpus Oral et Mail implique un fort

taux d'erreurs lors de l'analyse des relations dans ces corpus. En outre, les phrases longues et complexes dans certaines œuvres littéraires, du XVIIIe ou du XIXe siècle, complexifient la tâche d'analyse. Notons aussi que lors des évaluations, tant sur le corpus de développement que sur celui de test, si Cordial est arrivé second en analyse de relations, par contre le système classé premier n'était pas le même.

Les résultats par relation sont les suivants :

	Précision	**Rappel**	**F-mesure**		**Précision**	**Rappel**	**F-mesure**
SUJ_V	0,83	0,85	0,84	MOD_N	0,75	0,71	0,73
AUX_V	0,97	0,96	0,97	MOD_A	0,72	0,49	0,58
COD_V	0,68	0,66	0,67	MOD_R	0,43	0,57	0,49
CPL_V	0,57	0,63	0,60	MOD_P	0,00	0,00	0,00
MOD_V	0,61	0,46	0,52	COORD	0,55	0,46	0,50
COMP	0,65	0,53	0,58	APPOS	0,08	0,10	0,09
ATB_SO	0,52	0,52	0,52	JUXT	0,18	0,14	0,16

Figure 3 : résultats par relation de la pré-campagne Passage (arrondis à 2 décimales)

Par rapport aux autres participants de cette pré-évaluation, notre classement pour les différentes relations est le suivant :

- 1er pour Suj-V, Aux-V, Comp, Mod-A et Coord
- 2nd pour Cod-V, Cpl-V, Mod-V, Atb-So, Mod-N et Appos
- 3e pour Mod-R et Juxt

On notera que pour la relation Sujet-Verbe, qui demeure la plus importante tant pour la correction grammaticale que pour l'analyse sémantique de la phrase, notre score est excellent, d'autant que le meilleur de nos concurrents sur cette relation obtient 0,785583 de f-mesure. Pour la relation COD-verbe notre score est un peu inférieur au premier système qui obtient 0,687386 et pour la relation attribut du sujet, notre score est un peu plus éloigné du premier de nos concurrents qui obtient 0,595890. Si l'on prend en compte les relations les plus importantes (sujet-verbe, auxiliaire-verbe, attribut du sujet, cod-verbe), notre système a les meilleurs résultats globaux.

5 Conclusion

Toutes les bases théoriques et conceptuelles de cet analyseur ont été posées en 1990 et 1991. Près de vingt ans plus tard, elles ne semblent pas devoir être remises en cause et les développements récents effectués sur la langue anglaise nous ont montré que ces postulats pouvant s'appliquer à d'autres langues avec des résultats comparables, même si les ressources linguistiques créées pour l'anglais ne sont pas encore au niveau de celles créées pour le français.

Cordial est un analyseur à approche essentiellement statistique et probabiliste, même s'il utilise quelques règles pour la désambiguïsation grammaticale et surtout de très nombreuses règles pour la correction grammaticale. Les règles les plus importantes sont souvent des règles implicites, comme le fait qu'à un verbe ne saurait correspondre qu'un seul attribut ou un seul COD au maximum. Ces règles sont de fait induites par le code, qui n'accepte pas d'étiqueter plus d'un attribut ou d'un COD par verbe !

Cordial est un analyseur syntaxique d'une grande robustesse et d'une extrême rapidité, deux qualités capitales pour une exploitation industrielle. À la base de nombreux composants linguistiques, il constitue une brique de base de haut niveau, autorisant des traitements sémantiques en profondeur. Ses résultats, lors des récentes évaluations, montrent clairement sa supériorité en détection des constituants et en détection des relations sémantiquement fortes. Il reste toutefois perfectible et nous travaillons constamment à l'améliorer.

Références

BENZITOUN C., VERONIS J. (2005) « Problèmes d'annotation d'un corpus oral dans le cadre de la campagne EASY », *Actes des Ateliers de la 12e Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2005)*.

DE LA CLERGERIE É. (2007) « PASSAGE, Produire des Annotations Syntaxiques à Grande Échelle », *Grand Colloque STIC 2007*.

DE LA CLERGERIE É., HAMON T., MOSTEFA D., AYACHE C., PAROUBEK P., VILNAT A. (2008), « PASSAGE: from French Parser Evaluation to Large Sized Treebank » *Proceedings of LREC, Marrakech, 28-30 mai 2008*

HAMON O., POPESCU-BELIS A., CHOUKRI K., DABBADIE M., HARTLEY A., MUSTAFA EL HADI W., RAJMAN M., TIMIMI I. (2006) « CESTA: First Conclusions of the Technolangue MT Evaluation Campaign », *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

LAURENT D., NEGRE S. (2006). Cordial, le TAL et les aides à la rédaction. *Journées de l'ATALA, Paris, 3 juin 2006*.

MAKHOUL J., KUBALA F., SCHWARTZ R. AND WEISCHEDEL R. (1999) « Performance measures for information extraction », *Proceedings of DARPA Broadcast News Workshop*.

PAROUBEK P., ROBBA I., VILNAT A., POUILLOT L.-G. (2005) « Easy : Campagne d'évaluation des analyseurs syntaxiques », *Actes des Ateliers de la 12e Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN 2005)*.

PAROUBEK P, ROBBA I, VILNAT A., AYACHE C. (2006) « Data, Annotations and Measures in EASY - the Evaluation Campaign for Parsers of French », In proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006).

PAROUBEK P., VINAT A., ROBBA I., AYACHE C. (2007). Les résultats de la campagne Easy d'évaluation des analyseurs syntaxiques du français. *TALN Toulouse*

PAROUBEK P., ROBBA I., VINAT A. (2008). EASY : La campagne d'évaluation des analyseurs syntaxiques.in "L'évaluation des technologies de traitement de la langue", IC2, Hermès, Lavoisier, 2008, p. 117-140.

VERONIS J. (1998). Annotation automatique de corpus : état de la technique, *Colloque International "Questions de méthode dans la linguistique de corpus"* Perpignan (France), p. 7-9.