Apport des cooccurrences à la correction et à l'analyse syntaxique

Dominique Laurent, Sophie Nègre, Patrick Séguéla (1)

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

Résumé : Le correcteur grammatical Cordial utilise depuis de nombreuses années les cooccurrences pour la désambiguïsation sémantique. Un dictionnaire de cooccurrences ayant été constitué pour les utilisateurs du logiciel de correction et d'aides à la rédaction, la grande richesse de ce dictionnaire a incité à l'utiliser intensivement pour la correction, spécialement des homonymes et paronymes. Les résultats obtenus sont spectaculaires sur ces types d'erreurs mais la prise en compte des cooccurrences a également été utilisée avec profit pour la pure correction orthographique et pour le rattachement des groupes en analyse syntaxique.

Abstract : For many years, the spellchecker named Cordial has been using cooccurrences for semantic disambiguation. Since a dictionary of co-occurrences had been established for users of the spellchecker and of the writing aid, the richness of this dictionary led us to use it intensively for the correction, especially for homonyms and paronyms. The results are impressive on this kind of errors but taking into account the cooccurrences proved to be very profitable for pure spellchecking and for the attachment of groups in syntactic parsing.

Mots-clés : cooccurrences, collocations, correction grammaticale
Keywords: collocation, grammar checking

1 Introduction

1.1 Cordial

Cordial (acronyme de CORrecteur D'Imprécisions et Analyseur Lexico-sémantique) est un logiciel de correction grammaticale basé sur un analyseur syntaxique robuste. Développé au début des années 90 mais constamment maintenu et enrichi depuis, Cordial est commercialisé en tant que logiciel de correction et d'aide à la rédaction depuis 1995. il est également diffusé en OEM et a, dans ce cadre, équipé de nombreuses années la suite Office sur Windows et Mac. Une version spécifique, Cordial Analyseur, est destinée aux laboratoires de recherche.

1.2 Corpus d'extraction

À l'origine, les tables statistiques utilisées par Cordial, tant pour la désambiguïsation grammaticale que pour établir le contexte en désambiguïsation sémantique, avaient été établies à partir d'un corpus d'environ 150 millions de mots. Il y a une dizaine d'années, nous avons porté ce corpus à 500 millions de mots puis, en 2006, à 1,2 milliard de mots (très exactement 1 202 540 979 mots). Ce corpus étant à la base de notre dictionnaire de cooccurrences, le tableau ci-dessous fournit quelques indications sur ses différentes composantes :

Type de corpus	% relatif
Corpus administratif et juridique	14 %
Corpus encyclopédique (Wikipedia, Encyclopaedia Universalis)	27 %
Corpus Internet (sites : 43 %, news et blogs : 57 %)	7 %
Corpus journalistique (AFP, Le Monde, Le Monde diplomatique, l'Essor)	27 %
Corpus littéraire (4107 ouvrages, dont 45% de moins de 20 ans)	19 %
Corpus technique (893 ouvrages : dont médical (23 %), informatique: 9%)	6 %

Figure 1. Répartition par genres du corpus de 1,2 milliard de mots

Ce corpus est très récent : 79 % du corpus a moins de vingt ans et 87 % moins d'un siècle. Le corpus est réellement francophone : 85,8% des textes sont d'origine française, 9,1 % d'origine canadienne (Hansard, textes littéraires), 3,4% d'origine suisse, 1% d'origine belge et 0,7% d'origine africaine. Ce qui traduit une légère sur-représentation canadienne, une nette sur-représentation suisse et une sous-représentation belge et surtout africaine.

2 Dictionnaire de cooccurrences

Le dictionnaire de cooccurrences de Cordial a été constitué à partir du corpus décrit ci-dessus. L'ensemble des textes ont été analysés syntaxiquement et les cooccurrences de l'ensemble des mots et syntagmes ont été stockées. Le processus global a demandé environ 20 heures machine sur un Pentium 3 GHz. En fin d'analyse, un tri a permis de ne conserver que les mots et syntagmes dont les effectifs globaux et de combinaisons étaient suffisants pour être conservés, c'est-à-dire lorsqu'au moins une cooccurrence avait un effectif suffisant vis-à-vis de la fréquence du mot dans le corpus. Ainsi l'adjectif "intercités" figurait 315 fois dans le corpus mais 4 fois précédé de "réseaux" et 3 fois précédé de "liaisons". Les fréquences relatives de ces

cooccurrences sont suffisamment fortes, relativement à la fréquence simple du mot pour que ces deux cooccurrences et l'entrée soient conservées. De la même façon, pour le syntagme "*aller au lit*", qui apparaît 208 fois dans le corpus, la seule cooccurrence conservée sera "*heure d'*" pour la relation "verbe comme complément de nom", car cette cooccurrence apparaît 9 fois. La sélection des cooccurrences conservées s'est effectuée avec une prise en compte de la fréquence du mot cooccurrent, sachant que les auxiliaires et quelques mots très fréquents n'étaient pris en compte que si la fréquence de la cooccurrence était exceptionnelle.

Au final, le dictionnaire comporte 48 130 entrées, dont 32 307 mots simples et 15 823 syntagmes. L'ensemble de ces entrées pointe vers 1 150 000 combinaisons. Ce total était beaucoup plus important en fin d'analyse (plus de 37 millions de combinaisons) mais nous avons supprimé automatiquement toutes les cooccurrences n'ayant pas une fréquence significative, comparativement à la fréquence de chacun des constituants. En moyenne une entrée possède donc 24 cooccurrences, la fréquence moyenne des cooccurrences étant de 7.

Aucune correction manuelle n'a été effectuée sur les résultats de cette analyse, ce qui signifie que les erreurs figurant dans les résultats sont directement imputables à notre analyseur, ou parfois au programme de remise en forme des résultats qui, par exemple, peut parfois associer une forme masculine et une forme féminine de manière erronée. Nous avons utilisé le terme de "cooccurrences" mais il s'agit ici de mots ou de syntagmes en relation syntaxique. Le tableau ci-dessous répertorie l'ensemble des relations relevées :

Noms :
Nom + adjectifs épithètes antéposés ("graves abus", "nombreux abus"...)
Nom + adjectifs épithètes postposés ("abus sexuel", "abus possibles"...)
Nom + noms en complément ("abus de biens sociaux", "abus de confiance"...)
Nom + verbes en complément ("absurdité de vouloir prouver", "absurdité de croire"...)
Nom + noms en association ("abus et fraude", "abus et dépendance"
Nom + noms en alternative ("abus ou omission"...)
Nom + noms dont il est complément ("victimes d'abus", "cas d'abus"...)
Nom + adjectifs dont il est complément ("coupables d'abus", "susceptible d'abus"...)
Nom en sujet ("l'abus existe", "un abus se produit"...)
Nom en complément d'objet direct ("évite l'abus", "dénonce l'abus"...)
Nom en attribut ("paraît une absurdité", "semble une absurdité"...)
Nom en autre complément ("donne lieu à des abus", "ouvre la porte à l'abus"...)

Adjectif :
Adjectif nom en complément ("absent du procès", "absent des débats"...)
Adjectif verbe en complément ("aberrant de voir", "aberrant de débattre"...)
Adjectif adverbe modificateur ("souvent absent", "totalement absent"...)
Adjectif épithète en antéposition ("abjecte pauvreté", "abjecte créature"
Adjectif épithète en postposition ("air absent", "regard absent"...)
Adjectif attribut ("resté absent", "semblait absent"...)

Verbe :
Verbe adverbe modificateur ("s'abaisser encore", "s'abaisser lentement"...)
Verbe avec verbe en complément direct ("adorer faire", "adorer jouer"...)
Verbe avec verbe en complément indirect ("s'abaisser à demander", "s'abaisser à faire"...)
Verbe comme complément de nom ("décision d'abaisser", "intention d'abandonner"...)
Verbe comme complément d'adjectif ("prêt à s'abandonner", "possible de s'abonner"…)
Verbe ayant pour sujet ("la température s'abaisse", "les paupières s'abaissent"...)
Verbe ayant pour complément d'objet direct ("abaisse le taux", "abaisse le regard"...)

Verbe ayant pour attribut ("accomplit seule"...)
Verbe ayant pour autre complément (abaisse à x francs", "abaisse d'un quart"...)

Adverbe :
Adverbe modifiant l'adjectif ("abominablement faux", "abominablement ivre"...)
Adverbe modifiant le verbe ("souffre abominablement", "griser abominablement"...)
Adverbe modifiant l'adverbe ("absolument plus", "absolument d'accord"...)
Adverbe sujet ("beaucoup sont", "beaucoup restent"...)
Adverbe complément d'objet direct ("avoir beaucoup" "tenter beaucoup"...)

Au sens littéral, certaines des "cooccurrences" relevées ici sont spatialement éloignées puisque la distance, par exemple, entre le sujet et le verbe peut parfois dépasser les dix mots. Ce dictionnaire est accessible pour l'utilisateur dans le logiciel, sous l'item "Combinaisons de mots", en couplage avec un dictionnaire d'exemples fournissant des contextes pour chacune des cooccurrences. Il permet donc à l'utilisateur d'explorer la combinatoire lexicale, chaque cooccurrence fonctionnant comme un lien hypertexte vers le mot cooccurrent. L'importance donnée aux syntagmes (plus de 30 % des entrées du dictionnaire) est très spécifique de notre dictionnaire. Elle permet de fournir des cooccurrences sur des groupes de mots (par exemple "*abandon du projet de centrale nucléaire*") qui sont absents de tous les autres dictionnaires de cooccurrences, informatiques ou non.

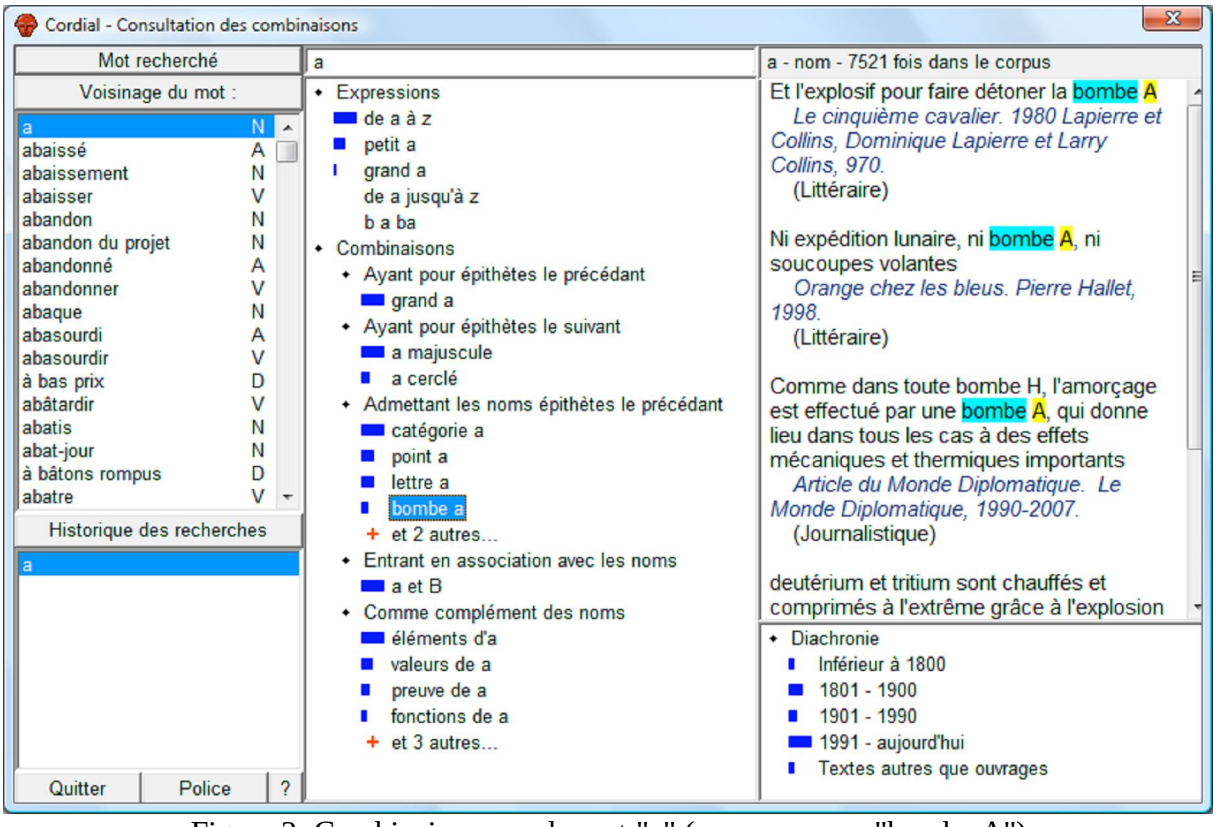

Figure 2. Combinaisons sur le mot "a" (cooccurrence "bombe A")

On notera que le dictionnaire fournit des indications diachroniques qui n'ont guère d'intérêt pour le mot "a" mais qui prennent toute leur valeur sur des mots sémantiquement plus "chargés". Ainsi le mot "âme" très employé jusqu'à la Belle Époque a quasiment disparu et le mot "amour" disparaît lui aussi peu à peu bien que plus lentement ! De la même façon, "homme" a une fréquence qui s'amoindrit, de même que "homme de coeur" mais "homme à

abattre" est considérablement plus fréquent qu'il y a un siècle... Quand à l'"homme de l'année", c'est une création très récente puisque toutes les occurrences ont moins de 20 ans !

3 Apport des cooccurrences en correction

Utiliser les cooccurrences en correction et en analyse syntaxique suppose de disposer d'un accès ultrarapide à cette ressource, faute de quoi l'analyse et la correction en seraient très ralenties. Pratiquement, cela signifie que le dictionnaire doit être chargé en mémoire. Sous sa forme texte, le fichier des cooccurrences occupe environ 35 Mo. Transcodé, il occupe environ 21 Mo, taille raisonnable pour un chargement en mémoire, sachant que la plupart des PC disposent d'au moins 1 Go de mémoire vive. Pour chacune des entrées, la fréquence globale du mot dans le corpus est conservée puis, pour chacune des cooccurrences, son type et sa fréquence. Cordial utilise ces cooccurrences pour deux types de corrections :

3.1 Correction orthographique

Rappelons que la correction orthographique consiste à fournir une ou plusieurs suggestions de remplacement lorsqu'un mot n'est pas trouvé dans les dictionnaires. Le dictionnaire de noms communs de Cordial contenant plus de 212 000 lemmes pour plus de 1,1 million de formes, lorsqu'un mot est inconnu, il est presque toujours erroné !

Fournir la bonne suggestion et si possible ne fournir qu'une suggestion constitue une vraie problématique, surtout si l'on désire effectuer une correction automatique avec un taux élevé de correction et une probabilité au moins égale à 99 %.

Sur un mot inconnu, Cordial recherche les suggestions par proximité lexicale (lettre oubliée, lettre ajoutée, interversion de lettres, oubli d'accent ou de majuscule, double lettre au lieu d'une simple et réciproquement, etc.), par proximité clavier (faute de frappe de proximité) et par phonétique (en utilisant un phonétiseur intégré). En couplant ces trois approches et en prenant en compte les fréquences des mots à suggérer dans le corpus, nous proposions une suggestion unique dans 67 % des cas et la meilleure suggestion en première position dans 83 % des cas, sur un corpus de 40 000 mots inconnus (issus de collations de textes depuis 20 ans) qui nous sert de corpus de test (nous utilisons comme corpus d'entraînement le "dictionnaire des fautes d'orthographe" de Michel Dansel, qui contient plus de 4000 mots fautifs). A une probabilité de 99 % (donc avec 1 % d'erreurs), nous pouvions remplacer automatiquement 43 % des mots inconnus.

La prise en compte des cooccurrences pour la correction orthographique a consisté à rechercher, pour chacune des suggestions, les cooccurrences éventuelles de ces mots avec les autres mots de la phrase. En associant la probabilité de cooccurrences de chaque suggestion aux probabilités de fréquence et de type de modification permettant d'aboutir du mot inconnu à la suggestion, nous obtenons les scores suivants :

Correction	sans cooccurrences	avec cooccurrences
une seule suggestion correcte	67 %	81 %
suggestion correcte en première position	83 %	94 %
correction automatique à 99 %	42 %	57 %

Figure 3. Apport des cooccurrences à la correction orthographique

L'apport des cooccurrences est particulièrement important pour le classement des suggestions et leur différenciation, il est moins important pour la correction automatique car la différence de probabilité entre la première suggestion et la seconde suggestion (souvent non affichée) doit être très élevée pour effectuer une correction automatiquement. Sans surprise, on notera que plus le mot est court, plus les suggestions sont nombreuses et plus il est difficile de déterminer quelle est la bonne suggestion. Une faiblesse résiduelle de l'algorithme de prise en compte des cooccurrences pour la correction orthographique provient de l'absence de mots grammaticaux (déterminants, pronoms, etc.) qui n'ont donc aucune cooccurrence lorsqu'ils figurent dans la liste des suggestions.

3.2 Correction grammaticale

La correction grammaticale consiste à remplacer un mot figurant dans les dictionnaires par un autre mot figurant dans les dictionnaires, généralement à la suite d'une faute d'accord. De fait, il existe d'autres types de fautes et, en particulier, les fautes d'homophonie (son/sont, tache/tâche...), d'homographie (le/la poste, le/la voile...), de paronymie (baiser/baisser, conjecture/conjoncture...). Ces fautes sont mal corrigées par la plupart des correcteurs car elles nécessitent une réelle analyse sémantique et contextuelle. Grâce à notre analyse sémantique, nous corrigions environ 41 % des fautes d'homonymie, 78 % des fautes d'homographie et 26 % des fautes de paronymie, sur un corpus de test de 1209 erreurs (issues de textes utilisateurs, de pages Web et du Canard Enchaîné) pour ces trois types confondus. Nous utilisons pour corpus d'entraînement un ensemble de 7359 fautes d'homonymie et de paronymie (les fautes d'homonymie sont issues de l'ouvrage de Jean Camion).

Les cooccurrences ne peuvent malheureusement pas être utilisées pour l'homographie car notre extraction automatique de cooccurrences n'a pas distingué entre les formes masculines et féminines des homographes (distinction souvent périlleuse faute de déterminant clair). Par contre, elles ont été largement prises en compte pour les homonymes et paronymes. Pour tester une éventuelle homonymie ou paronymie, Cordial utilise des marqueurs placés automatiquement dans nos bases grammaticales lorsque le mot figure dans notre dictionnaire d'homonymes et de paronymes. Ce dictionnaire contient actuellement 1907 homonymes et 815 paronymes, effectifs assez réduits car nous n'avons pas recherché l'exhaustivité, sachant que certains homonymes ou paronymes sont si rares qu'il serait très curieux qu'un utilisateur les emploie par erreur.

La prise en compte des cooccurrences consiste pour ces homonymes et paronymes marqués à comparer les contextes syntaxiques du mot et de son homonyme dans le dictionnaire. Comme pour la correction orthographique, les relations syntaxiques n'ont pas le même poids, les épithètes pesant plus lourd que les relation S-V V-A ou V-COD, qui pèsent elles-mêmes plus lourd que les autres relations. Les coefficients des relations ont été déterminés par apprentissage avec un corpus d'entraînement, ce sont malgré tout les fréquences relatives des cooccurrences qui pèsent le plus dans la décision.

Correction	sans cooccurrences	avec cooccurrences
d'homonymie	41 %	74 %
d'homographie	78 %	78 %
de paronymie	26 %	53 %

Figure 4. Apport des cooccurrence à la correction d'homonymie, homographie et paronymie

L'apport des cooccurrences à la correction grammaticale ne se limite pas aux homonymes et paronymes. Elle concerne également :

- l'affectation des adjectifs finaux de groupes nominaux prépositionnels ("*un pneu de la voiture crevé*" ou "*un pneu de la voiture crevée*"). Même le réseau sémantique utilisé par Cordial ne permet pas de déterminer si "*crevé(e)*" se rapporte à "*pneu*" ou à "*voiture*", s'agissant de deux objets concrets même si le second est animé et susceptible de transfert métonymique ("*je suis garé*" ou "*j'ai crevé*").

- le nombre de groupe nominaux prépositionnels ("*compagnon de voyage*" ou "*compagnon de voyages*"). Dans ce domaine, comme pour les homonymes et les paronymes, les correcteurs sont à la peine et les utilisateurs souvent hésitants. Dans de nombreux cas, le singulier et le pluriel sont admissibles mais on imagine mal "*un coup de balais*" ou "*un lobby de chasseur*").

- la position antéposée ou postposée de l'adjectif ("*de multiples exactions*" ou "*des exactions multiples*"). Il s'agit là plus d'une correction stylistique que d'une correction grammaticale. Ce type d'erreur est en fait assez rare, sauf pour des personnes dont le français n'est pas la première langue. Une utilisation antéposée au lieu de postposée peut aussi dénoter une volonté de surprendre ou de mettre l'accent sur l'adjectif comme dans "*un sec coup*".

4 Apport des cooccurrences en analyse syntaxique

Le dictionnaire de cooccurrences est également utilisé par Cordial en analyse syntaxique, comme outil d'aide au rattachement. En analyse syntaxique, les problèmes de rattachement sont multiples et complexes. L'utilisation des cooccurrences en correction a déjà abordé quelques cas de rattachement (de l'adjectif final à un des groupes nominaux, d'un groupe nominal à un autre) mais le correcteur ne s'intéresse qu'aux cas pouvant générer une faute de grammaire. Ainsi le module de correction traitera le cas du rattachement de "précédent" dans le groupe nominal "le document de référence précédent", afin de déterminer si "précédent" réfère à "document" ou à "référence" mais il ne traitera pas le cas de "page de référence précédente" puisque, dans ce cas, les deux rattachements possibles ont le même genre (féminin) et le même nombre (singulier).

4.1 Rattachement des adjectifs

En dehors du cas des adjectifs placés en fin de groupe nominal prépositionnel et pouvant s'accorder soit avec ce groupe nominal soit avec l'un des groupes nominaux précédents ("*le système de récupération de la chaleur installé*"), le problème du rattachement se pose également avec les adjectifs en position d'épithète d'un groupe nominal ou d'attribut du sujet ("*il restait dans la cour assis*", ou encore "*dans la chambre, patient, il attendait le sommeil*").

Dans le premier cas, un indice de probabilité de cooccurrence est établi entre l'adjectif final et chacun des noms. Ainsi dans l'exemple supra ("*le système de récupération de la chaleur installé*"), la cooccurrence de "*installé* " sera recherchée avec "*système*", "avec "*récupération*" et avec "*chaleur*". En fait, aucune cooccurrence n'est trouvée en relation épithète mais, comme "*installé*" est un participe passé, les cooccurrences en relation V-COD seront recherchées pour "*installer le système*", "*installer la récupération*", "*installer la chaleur*". En l'occurrence, la première relation ("*installer le système*") figure 283 fois contre 0 pour les deux autres

cooccurrences possibles. Notons ici que la relation "*installer le système*" est recherchée à l'entrée "installer" mais également à l'entrée "système". Car le dictionnaire ne constitue qu'une sélection des cooccurrences trouvées. Seules ont été conservées celles qui avaient un seuil de probabilité suffisant, seuil lié à la fréquence du mot ou du syntagme ainsi qu'à la fréquence relative de la relation. Les cooccurrences étant réciproques, les mêmes totaux seront trouvés mais il peut arriver qu'un seul total figure (parce que l'un des mots avait beaucoup plus de cooccurrences et que la cooccurrence en question a été élaguée) et, de toutes façons, ces totaux entrent dans le calcul de probabilité qui suit, uniquement en proportion de la fréquence globale de l'entrée et du type de relation pour ce mot ou ce syntagme.

4.2 Rattachement des groupes nominaux

Le raccordement des groupes nominaux concerne les groupes nominaux prépositionnels qui suivent un groupe nominal et qui peuvent être soit en position de rattachement à ce groupe nominal, soit en position de complément indirect ou circonstanciel du verbe, comme dans les exemples suivants : "*il a pris la route de bon matin*", ou "*il change le dossier de place*". Dans le premier cas, la nature adverbiale du complément de temps incite à détacher le groupe "*de bon matin*" du groupe nominal qui le précède pour le rattacher au verbe "*prendre*". En fait, le problème ne se pose même pas pour Cordial qui agglomère le syntagme verbal "*prendre la route*" ! Dans le second cas, le rattachement de "*place*" au verbe "*changer*" ou au groupe nominal "*le dossier*" est moins triviale.

En recherchant dans le dictionnaire de cooccurrences la relation verbe-COI pour "*changer*", on trouve 71 fois "*changer de place*" alors que la cooccurrence "*dossier de place*" ne figure pas (contrairement à "*dossier du patient*" ou "*dossier de chaise*"). Le rattachement de "*place*" au verbe "*changer*" s'impose donc ici. D'autant que notre dictionnaire de syntagmes verbaux offre l'entrée "*changer de place*", ce qui augmente considérablement la probabilité du rattachement.

Nous effectuons aussi quelques rattachements de verbes (participes) et d'adverbes lorsqu'il y a doute important mais ces cas sont rares.

5 Perspectives

Nous envisageons de réanalyser un corpus plus étendu, d'au moins 2 milliards de mots, afin de créer un dictionnaire de cooccurrences plus complet. L'espace mémoire occupé pourrait aller jusqu'à 40 Mo sans que cela soit réellement pénalisant pour les systèmes actuels. De plus la meilleure qualité de l'analyseur devrait permettre d'extraire les relations avec une fiabilité supérieure, d'où une amélioration globale de la qualité des résultats. Si nous avons utilisé quelques "jokers" par exemple pour désigner un nombre ou un pourcentage, il pourrait être intéressant de les étendre, par exemple aux types d'entités nommées en position sujet ou COD.

L'exploitation des cooccurrences en analyse syntaxique pourrait également être complétée. Actuellement plusieurs relations ne sont pas testées via le dictionnaire, par exemple les rattachements des coordinations, qui pourraient permettre d'améliorer le regroupement de groupes nominaux en position sujet et peut-être même la découpe en propositions. En effet, actuellement, les traitements exposés supra, excepté pour la correction orthographique, sont tous effectués après la découpe en propositions, ce qui est parfois trop tardif.

Nous n'avons pas encore testé l'apport éventuel des cooccurrences sur la désambiguïsation grammaticale. Il y a pourtant là une piste intéressante d'amélioration, même si Cordial commet peu d'erreurs à ce niveau. L'amélioration n'est d'ailleurs pas certaine en ce domaine car les relations étant le résultat de l'analyse de Cordial, sans aucune reprise manuelle, les erreurs potentielles de typage grammatical figurent sans doute dans notre dictionnaire...

D'un point de vue de recherche, il faudrait évaluer en détail l'apport de chacun des traitements sur l'analyse syntaxique. Lors de notre participation à Passage[1], nous avons déjà relevé une nette amélioration mais sans l'avoir quantifiée précisément.

Vis-à-vis de l'ensemble des améliorations possibles, un paramètre important doit être maîtrisé: le temps d'analyse supplémentaire que représentent ces nouveaux traitements. Ajouter le dictionnaire de cooccurrences dans notre chaîne de traitement a fait un peu baisser nos performances puisque notre vitesse d'analyse syntaxique est passée d'environ 12 000 mots/seconde à 10 000 mots/seconde. Même si cette vitesse est encore très importante et autorise les déploiements industriels, y compris dans des contextes de milliards de pages du Web, éviter de trop dégrader les performances par des traitements peu utiles est une nécessité et conditionnera nos décisions futures.

6 Conclusion

Synapse Développement a développé un dictionnaire de cooccurrences à grande échelle puisqu'il s'appuie sur un corpus de 1,2.milliard de mots. Avec plus de 48 000 entrées, dont plus de 15 000 syntagmes, ce dictionnaire constitue un outil riche pour les linguistes et tous les utilisateurs intéressés par la combinatoire lexicale, mais il constitue également un outil performant d'amélioration de l'analyse syntaxique et de la correction, orthographique et grammaticale.

Les cooccurrences sont utilisées tant pour l'analyse syntaxique que pour la correction et elles permettent une importante amélioration, en particulier dans le domaine de l'ordonnancement des suggestions et de la correction automatique des fautes d'orthographe, ainsi que pour la correction des homonymes et paronymes. Étendre les champs d'utilisation de ce dictionnaire de cooccurrences et mesurer l'apport de chacun des appels à ce module constituent pour nous deux tâches prioritaires.

[1] cf. http://atoll.inria.fr/passage/articles.fr.html, et "L'analyseur syntaxique Cordial dans Passage", TALN 2009, Senlis, 24-26 juin 2009

Références

AUDIBERT L., (2003). Désambiguïsation lexicale automatique : sélection automatique d'indices. *Actes de TALN 2007, Toulouse, 5-18 juin 2007,* Tome II, p. 13-22.

BEAUCHESNE J., (2001). *Dictionnaire des cooccurrences.* Éditions Guérin, Montréal, Québec, 2001, 402 pages.

BOURIGAULT D., FABRE C., (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire 25, 2000, "Sémantique et Corpus,* p. 131-151.

CAMION J., (1986). *Dictionnaire des homonymes de la langue française.* Éditions Gachot, Paris, France, 1986, 684 pages.

CHAREST S., BRUNELLE É., FONTAINE J., PELLETIER B. (2007) « Élaboration automatique d'un dictionnaire de cooccurrences grand public», *Actes de TALN 2007, Toulouse, 5-18 juin 2007,* tome I, p. 283-292

DANSEL M., (1995). *Dictionnaire des fautes d'orthographe.* Éditions du Rocher, Monaco, 1995, 274 pages.

DUBREIL E., (2008). Collocations : Définitions et problématiques. *Texto !* janvier 2008, vol. XIII, n° 1, p. 32-38.

KILGARIFF A. TUGWELL D., (2001). WORD SKETCH : Extraction and Display of Significant Collocations for Lexicography. *Proc. Collocations workshop, ACL 2001, Toulouse,* p. 32-38.

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

TUTIN A. (2004). Pour une modélisation dynamique des collocations dans les textes. *Actes d'EURALEX, Lorient, 6-10 juillet 2004,* tome 1, p. 207-221.

VOLK M. (2000). Using the WWW to resolve PP attachment ambiguities. Proc. of Konvens-2000. Ilmenau.

VOLK M. (2001). Exploiting the WWW as a corpus to resolve PP attachment ambiguities. Proc. of Corpus Linguistics 2001. Lancaster.

ZINGLE H., BROBECK-ZINGLE M.-L. (2003). *Dictionnaire combinatoire du français.* La Maison du Dictionnaire, Paris, 2003, 1306 pages.