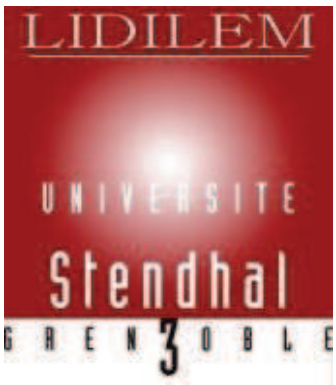
Prise en compte de variables syntaxiques et textuelles dans l'analyse sémantique distributionnelle automatique

Mickaël Galop



Prise en compte de variables syntaxiques et textuelles dans l'analyse sémantique distributionnelle automatique

Nom : GALOP
Prénom : Mickaël

UFR *LLASIC* DEPARTEMENT *SCIENCES DU LANGAGE*

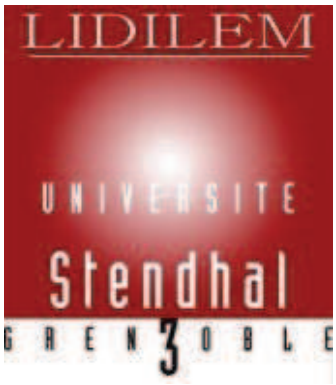
Mémoire de master *2 recherche* - 30 crédits – *Sciences du Langage*

Spécialité : *Modèles et traitements en industrie de la langue*

Parcours : *Traitements automatiques des langues écrites et parlées*

Sous la direction de *AGNES TUTIN*

Année universitaire 2010-2011



Prise en compte de variables syntaxiques et textuelles dans l'analyse sémantique distributionnelle automatique

Nom : GALOP
Prénom : Mickaël

UFR *LLASIC* DEPARTEMENT *SCIENCES DU LANGAGE*

Mémoire de master *2 recherche* - 30 crédits – *Sciences du Langage*

Spécialité : *Modèles et traitements en industrie de la langue*

Parcours : *Traitements automatiques des langues écrites et parlées*

Sous la direction de *AGNES TUTIN*

Année universitaire 2010-2011

Remerciements

Je tiens d'abord à remercier Agnès Tutin pour m'avoir proposé ce sujet, et pour m'avoir apporté ses connaissances et réflexions, qui m'ont guidé dans les choix que j'ai du faire. Je la remercie aussi pour son soutien durant toute la durée du mémoire.

Je remercie aussi Guillaume Jacquet, pour son aide, Caroline Hagege pour nous avoir aidé à utiliser XIP, et à travers eux Xerox pour nous avoir autorisé à utiliser leur analyseur syntaxique.

Je remercie aussi Olivier Kraif, particulièrement pour son aide au niveau algorithmique et programmation.

Enfin, je remercie Agnès Tutin, Guillaume Jacquet, Christelle Cavalla et Olivier Kraif pour avoir évalué les résultats de cette étude.

Sommaire

Introduction

Notre travail se place dans le cadre du traitement automatique des langues. Notre étude se concentrera principalement sur les liens entre la syntaxe et la sémantique, en étendant l'utilisation de l'analyse distributionnelle automatique. Nous ferons ainsi une étude sémantique à travers la syntaxe.

La distribution d'un mot contient l'ensemble des contextes dans lequel ce mot apparaît, ce que nous appellerons leur environnement. Cet environnement peut être de différentes sortes. Certaines études utilisent comme environnement une fenêtre de mots [Grefenstette, 1993], d'autres les relations syntaxiques [Harris, 1968]. Nous allons quant à nous étudier les mots dans un environnement lexico-syntaxique, en faisant varier cet environnement afin d'en observer les conséquences. Nous utiliserons les résultats d'une analyse syntaxique afin d'effectuer une analyse sémantique des mots des textes. L'environnement des mots sera donc composé de relations syntaxiques les liant à d'autres mots, et sera également composé de ces mots là. Prendre en compte la syntaxe permet de voir le lien entre deux mots qui ne sont pas voisins, qui sont séparés par un bout de phrase. Par exemples : *« Les prédateurs sont, quand ils chassent leurs proies, très agressifs ».* Dans cette phrase, si l'on n'intègre qu'une fenêtre limitée de mots, on ne peut pas relier *agressifs* à *prédateurs*.

Nous observerons également la position des mots dans la structure du document. Nous pensons que ce type de paramètre est pertinent pour l'analyse sémantique. Une phrase présente dans le résumé ou dans un titre est peut-être plus pertinente pour étudier la sémantique des mots qu'une phrase présente dans le contenu d'un paragraphe. Ceci a pour but de mettre en évidence un éventuel lien entre ces informations et le rapprochement sémantique des mots étudiés.

Nous pensons également que certaines relations syntaxiques sont plus pertinentes que d'autres au regard des informations qu'elles peuvent donner sur la sémantique des mots. Nous nous efforcerons de le vérifier.

Ce mémoire est effectué sous la tutelle d'Agnès Tutin, maître de conférences en Sciences du Langage à l'Université Stendhal et en interaction avec Guillaume Jacquet, ingénieur de recherche en Traitement Automatique des Langues Naturelles à Xerox RCE. Le travail présenté dans ce mémoire a donc été effectué au laboratoire LIDILEM, du l'Université Stendhal Grenoble 3.

Le LIDILEM concentre ses recherches autour de trois axes : les descriptions linguistiques, TAL, corpus ; la sociolinguistique et acquisition du langage ; et enfin la didactique des langues, recherches en ingénierie éducative. Le premier de ces axes se décompose en deux grands programmes : la syntaxe, sémantique, pragmatique et le Traitement Automatique des Langues. Nous travaillerons dans le cadre des deux programmes de l'axe 1.

Cette étude est très exploratoire, et a pour but de faire ressortir de nouveaux axes de recherche dans un domaine en plein essor. Les hypothèses que nous faisons ne se vérifieront peut-être pas, mais nous pensons qu'elles soulèveront quand même de nouvelles questions, qui pourront donner lieu à de nouvelles études. Nous allons présenter l'étude un peu plus en détails, puis expliquer les expériences que nous avons faites et les résultats obtenus.

Partie 1

Présentation de l'étude

Chapitre 1 - Présentation du domaine

Le Traitement Automatique des Langues, parfois appelé Traitement Automatique du Langage naturel, est un domaine rapprochant l'informatique et la linguistique. Il permet de créer des applications telles que les analyseurs syntaxiques, les traducteurs automatiques ou les correcteurs orthographiques.

Pour pouvoir recréer, reproduire informatiquement, il faut comprendre comment l'homme écrit. Ceci motive les études comme la nôtre.

Traditionnellement, ces logiciels analysent les phrases au niveau syntaxique, puis ensuite au niveau sémantique. Je pense que c'est d'ailleurs ce que font les êtres humains, quand ils lisent des phrases. L'analyse syntaxique est assez simple à effectuer, car elle repose sur un certain nombre de règles, souvent simples. L'analyse sémantique est plus complexe car elle fait apparaître différents problèmes, notamment ceux liés à la désambiguïsation. Nous chercherons à trouver dans quelle mesure nous pouvons utiliser les informations récupérées dans l'analyse syntaxique afin de créer des cliques de mots sémantiquement apparentés, c'est à dire un groupe de mots dont tous les membres sont reliés entre eux par une relation donnée, notre relation étant leur environnement lexico-syntaxique.

Les études dans ce domaine, comme celle-ci, aident à la construction de ressources textuelles. En effet, nous espérons pouvoir aider à construire des ressources fiables, et interrogeables. Les études dans le domaine ont également pour but de permettre la création de requêtes pour interroger ces ressources. On peut par exemple imaginer créer une requête interrogeant le corpus pour trouver les mots sémantiquement proches d'un mot donné. D'un point de vue plus immédiat, cette étude permettra de travailler sur le corpus Scientext, pour des applications didactiques.

Précisons que nous n'étudierons pas les textes phrases par phrases, mais dans leur ensemble, et même sur l'ensemble de notre corpus, car l'environnement lexico-syntaxique des mots sera enrichi par les relations autour de ces mots, relations trouvées dans différents textes.

Chapitre 2 – Etat de l'art

L'intérêt des linguistes pour la sémantique des mots n'est pas nouveau, et les méthodes linguistiques pour l'extraire ne le sont pas non plus. La première méthode qui a suscité notre intérêt fut proposée par un des premiers chercheurs à s'intéresser au sujet, Zelling.S. Harris [Harris, 1968]. Cette méthode, dite « distributionnelle », rapproche des mots sur la base de contextes lexicaux, formant ainsi des listes de « voisins » sémantiques. Des mots qui partagent les mêmes contextes vont ainsi appartenir à un même groupe. Par exemple, prenons *demander une étude* et *demander une analyse*. *étude* et *analyse* sont tous deux objet du verbe *demander*, qui fait donc partie de leur contexte commun. Cette méthode fut utilisée dans beaucoup de travaux par la suite.

Nous nous sommes intéressés à une seconde méthode, l'approche dite « par fenêtre » de Martin Phillips [Phillips, 1985]. Contrairement à l'approche « distributionnelle », l'approche « par fenêtre » ne s'intéresse pas à la syntaxe des mots mais utilise un étiquetage des mots. Le contexte d'un mot est alors simplement formé par les mots qui l'entourent.

Gregory Grefenstette [Grefenstette, 1993] mit au point une méthode pour évaluer les deux techniques précédemment citées. Cette évaluation, appliquée à un corpus de 4 méga octets, montre que l'approche syntaxique donne de meilleurs résultats sur les noms les plus fréquents du corpus alors que l'approche « par fenêtre » est meilleure lorsque l'on s'intéresse aux mots rares. Nous pensons que les mots les plus fréquents sont plus significatifs, ce qui nous pousse vers l'approche syntaxique.

Certaines études ont toutefois montré que l'on peut aussi classer les mots rares avec l'approche « distributionnelle ». D. Lin et P. Pantel [Lin, Pantel, 2001] ont présenté un algorithme, UNICON, capable de créer des clusters sémantiques, tout en apportant certains avantages par rapport aux travaux précédents. Il peut classer les mots dans un espace de grande dimension et aussi prendre en compte les mots souvent vus comme « inconnus ».

Cet algorithme fait appel à un autre algorithme : CLIMAX. CLIMAX s'exécute sur une liste d'éléments et sort une liste de clusters, qui sont similaires à des cliques sémantiques. Il utilise une matrice de similarités précédemment calculée. Pour chaque mot de la matrice, il garde les n mots les plus proches, n étant préalablement choisi par les chercheurs. Cela donne des clusters, des groupes de mots similaires. Ensuite, UNICON recherche les centroids de ces clusters, y applique plusieurs opérations, et sort à nouveau une liste de clusters.

Les centroids sont les mots les plus proches d'un maximum de mots du cluster. Ce sont donc les mots les plus représentatifs du cluster, utiles pour être l'étiquette du cluster.

Les tests de ces algorithmes ont été confiés à des juges. Les résultats sont plutôt concluants.

L'influence de la syntaxe sur la sémantique des mots a déjà été étudiée. D.Bourigault a utilisé un outil d'analyse distributionnelle, UPERY [Bourigault, 2002], basé sur l'analyseur syntaxique SYNTEX [Bourigault, Fabre, 2000], qui lui a permis d'avancer que certaines catégories syntaxiques créaient plus facilement un contexte. En effet, dans son étude, il montre que les adjectifs et les noms ont plus souvent un contexte que les adverbes et syntagmes nominaux. L'idée que certaines relations syntaxiques seraient plus pertinentes que d'autres est défendue par Agnès Tutin dans son article de 2007 [Tutin, 2007]. Son étude tend à montrer que les relations qui mettent en jeu des arguments souvent obligatoires seraient plus significatives. Ces observations ont motivé notre étude, dans laquelle nous essaierons de les vérifier, et d'aller plus loin dans leur exploitation.

Les travaux menés en linguistiques se basent le plus souvent sur des corpus spécialisés. Z. Harris pensait que la méthode « distributionnelle » échouerait si elle était utilisée sur des corpus non spécialisés. Cet échec serait causé par le grand nombre de mots polysémiques que l'on trouve dans les corpus généraux. Edith Galy et Didier Bourigault ont démontré que l'hypothèse de Z. Harrris était vérifiée. Il ont testé l'utilisation d'un corpus spécialisé dans un domaine d'un corpus général [Galy, Bourigault, 2005]. Ils ont obtenu une grande différence des résultats, qui étaient bien meilleurs avec le corpus spécialisé.

Nous allons utiliser un corpus spécialisé. Spécialisé à la fois sur le type de textes, il est composé d'écrits scientifiques ; mais aussi spécialisé sur le lexique, les textes étant scientifiques. Nous faisons donc l'hypothèse que la méthode « distributionnelle » présentera des résultats intéressants.

Chapitre 3 – Hypothèses et problématique

Nous faisons l'hypothèse que l'on peut rapprocher sémantiquement des mots qui ont le même contexte lexico-syntaxique. Nous pensons que l'environnement lexical des mots peut aider à créer un tel rapprochement. Notre idée est que deux mots, lemmatisés, étant souvent sujet ou objet des mêmes verbes, ou étant reliés par d'autres relations syntaxiques à de mêmes autres lemmes ont de grandes chances d'être reliés sémantiquement. La forme lemmatisée d'un mot est sa forme canonique, soit l'infinitif pour un verbe, le singulier pour un nom, et le masculin singulier pour les adjectifs. Par exemple les noms *études*, lemmatisé en *étude* et *recherches*, lemmatisé en *recherche* seront objets des mêmes verbes *effectuer*, *demander*, *montrer* ....

Nous faisons également l'hypothèse que certaines relations syntaxiques sont plus pertinentes que d'autres. La relation sujet-verbe peut être plus pertinente que les relations adjectivales, car les adjectifs ne servent souvent qu'à préciser le nom, on peut les enlever sans trop changer le sens de la phrase.

Enfin, nous faisons l'hypothèse que la position de la phrase dans la structure du document a une importance. En effet, un titre, ou le résumé, servant à la fois à présenter et à résumer ce qui va suivre, nous pouvons donc penser qu'ils expriment ce qu'il y a de plus important, et donc qu'ils sont particulièrement aptes à véhiculer le sens. Notons toutefois qu'il y a moins de relations syntaxiques dans les titres, et qu'elles sont moins diversifiées.

Nous allons donc étudier dans quelle mesure les différents types de relations syntaxiques influent sur la proximité sémantique entre les mots. Nous allons essayer de vérifier que, par exemple, on peut considérer la relation sujet comme plus significative que la relation objet, bien que Cécile Fabre mentionne que la relation sujet-verbe est moins pertinente que la relation verbe-objet [Fabre,2010]. Nous prendrons également en compte la position des mots dans la structure du document, c'est à dire le fait qu'ils se trouvent dans un titre, ou dans un paragraphe, ou autre part, pour vérifier notre autre hypothèse. Nous avons pour objectif d'utiliser ces informations pour créer des groupes de mots sémantiquement apparentés.

Partie 2

Préparation de l'expérimentation

Chapitre 4 – Le corpus

Scientext

D'un point de vue technique, nous utiliserons le corpus d'écrits scientifiques Scientext[1]. Ce corpus, mis gratuitement à la disposition des chercheurs et des étudiants, est composé d'un grand nombre de textes scientifiques. Ces textes sont pour certains écrits en langue anglaise, les autres sont écrits en français. Nous n'étudierons que les textes écrits en français.

Nous utiliserons un échantillon de 458 textes, que nous soumettrons à différents pré-traitements que nous détaillerons plus tard. Ce panel de textes, après avoir appliqué ces modifications, contiendra environ six millions de mots. Ce corpus est annoté afin de pouvoir classer les textes selon la discipline (sciences humaines, expérimentales ou de l'ingénieur) ainsi que le genre (thèses, articles, communications, Habilitations à Diriger des Recherche). Nous donnons dans la figure ci-dessous le nombre de textes, et de mots, présents pour chacune des disciplines du corpus. Le nombre de mots tient compte des pré-traitements que nous développons dans la prochaine section.

Discipline	Biologie	Economie	Electronique	Linguistique	Mécanique	Médecine
Nombre de textes	28	50	5	149	5	58
Nombre de mots	516819	365351	371718	1828195	200202	232996

Discipline	Psychologie	Sciences de l'éducation	Traitement Automatique des Langues
Nombre de textes	57	89	17
Nombre de mots	748471	1554565	572979

Figure 1 : Nombre de textes et de mots par discipline

Nous pouvons voir sur la figure 1 que les disciplines ne sont pas représentées équitablement, ce dont nous tiendrons compte dans notre étude. Nous voyons également que la discipline « électronique » et la discipline « économie » contiennent autant de mots alors que l'une englobe dix fois plus de textes que la seconde. Cela s'explique par le fait que les textes traitant de l'électronique sont des thèses, alors que ceux traitant d'économie sont des articles.

Nous utiliserons une version de ce corpus qui est annotée structurellement, dans un format XML. Nous aurons ainsi les informations sur la position des phrases dans la structure du document. Nous donnons en annexe[2] le début d'un fichier de Scientext au format XML.

1 http://scientext.msh-alpes.fr/

2 Annexe 1 : Un exemple de fichier de Scientext

Les méta-informations, comme par exemple le genre et la discipline, se situent dans l'entête du document. Le document est ensuite découpé en parties (introduction, sections, conclusion ...) qui sont balisées. A l'intérieur de ces parties un autre balisage permet d'identifier les titres et les paragraphes.

Partie	Introduction	Résumé	Développement	Conclusion	Notes	Titre
Nombre de textes	334	399	458	328	458	458
Nombre de mots	352466	61316	5260844	161299	340958	76541

Figure 2 : Nombre de textes et de mots par partie structurelle

Sans surprise, nous voyons sur la figure 2 que le nombre de mots présents dans les résumés et les titres est égale à environ 2% de l'ensemble des mots, soit 137857. C'est tout de même un nombre suffisamment élevé pour que l'on puisse faire une expérimentation qui n'utiliserait que les mots présent dans les résumés et les titres, ou que l'on utilise une pondération différente.

Ce formatage des textes est très approprié pour l'étude que nous entreprenons, vis à vis des objectifs cités ci-dessus, et facilitera un certaines nombres d'opérations que nous évoquons dans la suite.

Les pré-traitements

Avant d'analyser syntaxiquement le corpus Scientext, nous sentons la nécessité d'effectuer quelques traitements préalables qui ont pour but d'éviter quelques biais et de se prémunir contre le « bruit », à savoir du texte non pertinent. Nous commençons bien sûr par éliminer les méta-informations liées au format xml. Nous veillons évidement aussi à supprimer les occurrences de mots ou phrase écrites dans une autre langue que le français.

Nous supprimons également les tableaux, figures, exemples et mots clés, car ils ne contiennent pas de phrases. Mais nous posons aussi la question de la pertinence de certaines sections des textes. La bibliographie, par exemple, n'est pas pertinente à étudier, car comme les exemples, elle ne contient pas de phrases. Il en est de même pour les glossaires, les annexes et les abréviations. Enfin, nous enlevons les remerciements. Ceux-ci sont composés de « vraies phrases », mais qui n'ont pas de rapport avec le reste du texte. Il s'agit de ce que l'on appelle communément le « péritexte ». Nous avons choisi Scientext pour la spécialisation scientifique du corpus, et les remerciements sont indépendants du thème évoqué dans le texte. Maintenant que nous avons choisi un corpus, nous devons choisir un analyseur syntaxique à utiliser.

Chapitre 5 – L'analyseur syntaxique

XIP

Afin de choisir l'analyseur syntaxique que nous utiliserons, nous en avons étudié plusieurs : Connexor[1], SYNTEX, XIP[2]. Nous avons finalement choisi d'utiliser l'analyseur XIP, pour différentes raisons. Cet analyseur est déjà utilisé ou va l'être dans d'autres projets menés par le laboratoire LIDILEM. Nous pourrons donc nous appuyer sur les connaissances engrangées par les équipes qui l'ont déjà utilisé et sans doute apporter à notre tour de nouvelles connaissances. De plus, ce projet étant également en collaboration avec Mr Guillaume Jacquet, nous pourrons bénéficier de l'expertise des chercheurs de Xerox.

XIP, Xerox Incremental Parser, est une application qui permet d'analyser profondément ou superficiellement des textes découpés en phrases, paragraphes ou tels quels. Il traite les données, comme son nom l'indique, de manière incrémentale. Cela signifie que la sortie d'une étape de l'analyse servira d'entrée pour l'étape suivante. Cette modularité nous permettra de personnaliser XIP selon nos besoins, comme expliqué plus loin. Le fonctionnement de XIP est expliqué plus en détail par ses créateurs [Aït-Mokhtar, Chanod, Roux, 2001]. Nous avons été amené à créer de nouvelles relations de dépendance, qui ont été implémentées dans XIP. Nous avons eu le soutien de Salah Aït-Mokhtar, qui a de l'expérience dans la création de nouvelles dépendances.

Méta-règles

Nous commençons par simplifier les relations transitives, notamment les relations de coréférence et de coordination. Nous le faisons en créant des méta-règles. Elles portent sur n'importe quel type de relations syntaxiques. Prenons le bout de phrase *« la personne qui voit »*. XIP, comme sans doute d'autres analyseurs syntaxiques, voit deux relations : une relation sujet entre les mots *voit* et *qui*, et une relation de coréférence entre *personne* et *qui*. Nous pensons que la « vraie » relation, celle qui est légitime au plan sémantique, est la relation sujet entre *personne* et *voit*.

1 http://www.connexor.eu/

2 http://www.xrce.xerox.com/Research-Development/Document-Content-Laboratory/Parsing-Semantics/Robust-Parsing

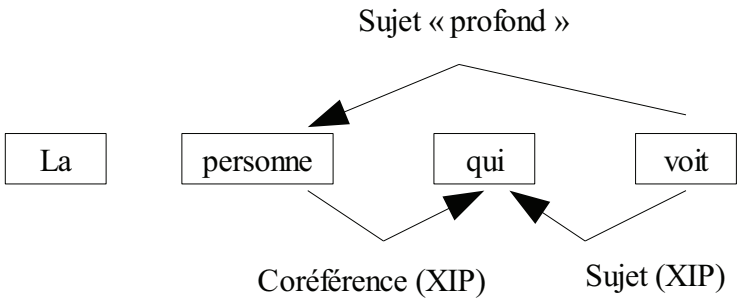


Figure 3 : Une méta-règle

Le raisonnement est le même pour la coordination. Dans la phrase *« Chouchou et Loulou mangent.»*, l'analyseur relie « Chouchou » à « mangent » et « Loulou » à « Chouchou », alors que de notre point de vue, « Loulou » doit être relié à « mangent ». Nous employons la même démarche que celle proposée par Caroline Hagège et Claude Roux [Hagège, Roux, 2003].

Modifications de règles

En outre, dans nos relations nous choisissons d'éclater la relation modifieur de XIP.

Dans le cas d'un modifieur avec préposition comme *« donner à Lulu »*, nous obtiendrons avec XIP deux relations : <vmod,donner,Lulu> et <prepobj,Lulu,à>. Ces deux relations sont trop vagues, et seront donc difficiles à utiliser pour notre objectif d'analyse sémantique. Nous trouvons donc intéressant de créer une relation <a_vmod,donner,Lulu>. Il en va de même pour les prépositions *de*, *sur*, *par*, *selon*, *en*, *pour*, *dans* et *après*.

Notre dernière modification concernera les formes passives. Dans une phrase du type *« il est soulevé par la tornade »*, *tornade* est un complément d'agent, alors que sur le plan sémantique c'est le sujet et *il*, vu comme le sujet, est en réalité l'objet.

Suppressions de règles

Nous supprimons les relations auxiliaires, ainsi que les relations liées à l'auxiliaire *aller* et aux modaux *pouvoir* et *devoir*. Nous pensons en effet que dans une phrase telle que *« il devra marcher »*, l'information sémantique importante est portée par *il [marchera]*, le modal n'est pas pertinent pour cette étude. Nous faisons également l'hypothèse que la relation « déterminant » n'est pas pertinente. Nous pourrions refaire les expériences qui suivront en intégrant cette relation pour valider notre hypothèse, mais ce ne sera pas l'objet de cette étude.

Pour finir, nous supprimons les relations unaires, telle que la négation, car nos distributions seront basées sur les contextes lexico-syntaxiques, ce qui sous entend des relations entre un mot et d'autres mots et exclut les relations unaires.

Nous récapitulons l'ensemble des modifications de règles dans le tableau ci-dessous, dans la page qui suit et se lit au format paysage.

Modification	Phrases (ou partie de)	XIP Avant	XIP Après
Coréférence	*La personne qui voit*	Subj(voit,qui) ; Coref(personne,qui)	Subj(voit,personne)
Coordination	*Le chien et le chat mangent*	Coorditem(chien,chat) ; Subj(chien,mangent)	Subj(chien,mangent) ; Subj(chat,mangent)
Modifieur adjectival	*C'est une maison bleue*	Nmod(maison,bleue) ; bleue est adjectif	Adj_nmod(maison,bleue)
Apposition nominale	*Le prix Nobel*	Nmod(prix,nobel) ; Nobel est un nom	N_mod(prix,nobel)
Complément du nom en de, à, sur, par, selon, en, pour, dans	*L'accord de l'entreprise*	Nmod(accord,entreprise) ; Prepobj (entreprise,de)	De_nmod(accord,entreprise)
Modifieur adverbiale du verbe	*Il a rapidement obtenu*	Vmod(obtenu,rapidement) ; rapidement est adverbe	Advmod(obtenu,rapidement)
Complément du verbe en de, à, sur, par, selon, dans, après, en ,pour	*Il décide de travailler*	Vmod(décide,travailler) ; Prepobj(travailler,de)	De_vmod(décide,travailler)
Nom-attribut adjectivale	*La fille est gentille*	Attr(est,gentille) ; Subj(est,fille)	Adj_nmod(gentille,fille)
Subordonnée	*Je sais qu'il marche*	Vmod(sais,marche)	Subord(sais,marche)
Les passifs	*Il est présenté par l'ingénieur*	Deepsubj(présenté,ingénieur)	Subj(présenté,ingénieur)

Figure 4 : Récapitulatif des modifications de XIP

Chapitre 6 – Les sorties

Format TXS

Il nous semble important d'effectuer notre étude de façon modulaire. Notre souhait est que si d'autres équipes souhaitent dans le futur reprendre cette étude, avec d'autres hypothèses, ou pour l'étendre, elles n'aient pas à refaire tout le travail de prérequis. C'est pourquoi nous décomposons nos traitements en différentes étapes, comme indiqué dans la figure 5.

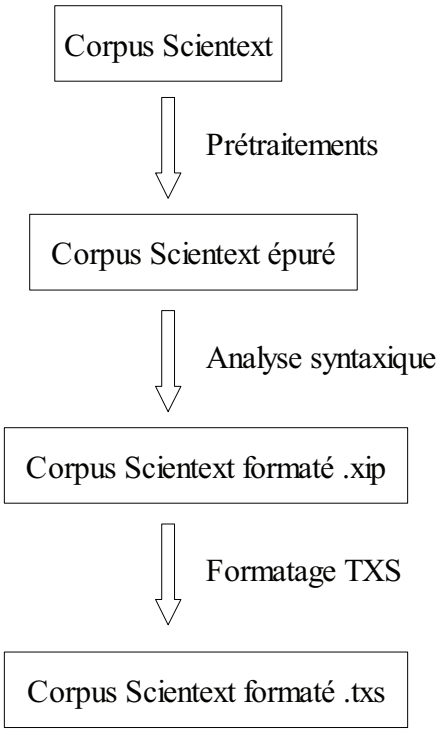


Figure 5 : Les étapes de transformation du corpus Scientext

Les premières étapes étaient les prétraitements et l'utilisation de XIP. Après avoir analysé le corpus Scientext, nous obtenons un fichier de sortie propre à XIP. Ce fichier n'est pas directement exploitable car il est peu lisible. Nous le transformons donc en un fichier au format TXS, format utilisé par d'autres équipes de recherches qui exploitent le même type d'entrées. Ce format, qui est une version compacte d'un format similaire, cesAna, a été créé et est utilisé par Olivier Kraif et ses collègues, qui sont des membres du LIDILEM.

Nous utilisons ce format car il permet de bien afficher les informations syntaxiques des mots ainsi que leurs liens de dépendance avec les autres mots de la phrase. Nous l'utilisons aussi par souci de standardiser les formats d'entrées-sorties utilisés par les différentes équipes du laboratoire.

Nous donnons un exemple de sortie sous ce format en annexe[1]. Nous pouvons maintenant constituer une liste de relations « mot1, mot2, relation ». En voici quelques unes.

Mot 1	Type 1	Mot 2	Type 2	Relation
développement	NOUN	lecture	NOUN	DE_NMOD
valeur	NOUN	impact	NOUN	SUR_NMOD_inv
structure	NOUN	linguistique	ADJ	ADJ_NMOD
travail	NOUN	terrain	NOUN	DE_NMOD
fonction	NOUN	linguistique	ADJ	ADJ_NMOD

Figure 6 : Exemples de relations extraites par notre analyse syntaxique

Mots à analyser

Nous sauvegardons cette liste dans un format facilement utilisable par quiconque serait intéressé.

Nous voulons étudier les relations qui mettent en jeu les noms, car nous pensons qu'elles sont particulièrement porteuses d'informations sémantiques. Mais tous les noms ne sont pas aussi pertinents. Nous sélectionnons une liste de noms sur deux critères liés à une problématique qui intéresse l'équipe : le lexique transdisciplinaire [Tutin, 2007].

Le premier est la fréquence d'apparition du nom dans le corpus, en tenant compte du fait que les disciplines des fichiers ne sont pas équitablement réparties. Pour être précis, nous avons choisi de garder un nom si il est présent avec une fréquence supérieure à 1/5000ème de l'ensemble des mots de la discipline. Par exemple, le corpus contient, après prétraitements, 1828195 mots dans des articles, HDR ou thèses de linguistique. Ne peuvent être retenus que les noms présents au moins 1828195/5000 = 366 fois dans les études linguistiques. Nous avons choisi ce facteur 1/5000 pour qu'il y ait assez de candidats, après avoir fait plusieurs tests.

Le deuxième critère est la transdisciplinarité des mots. Agnès Tutin définit le lexique transdisciplinaire des écrits scientifiques comme étant le lexique partagé par la communauté scientifique. Il renvoie aux concepts mis en œuvre dans l'activité scientifique [Tutin, 2007].

1 Annexe 2 : Exemple de contenu d'un fichier TXS

Il contiendra donc des mots tels que *expérience*, *hypothèse* et *résultat*. Nous décidons d'une mesure de transdisciplinarité. Nous ne gardons que les mots présents, en nombre d'occurrences suffisant, dans au moins 6 des 9 disciplines du corpus. Nous obtenons une liste de 91 noms. Parmi ces noms, nous en remarquons 4 qui nous posent problème.

Le mot *autre* est plutôt employé en tant qu'adjectif, nous le supprimons donc de la liste.

Le mot *effet* apparaît souvent dans des expressions comme *en effet*, c'est pourquoi nous le retirons. Il en va de même pour le mot *mise* qui apparaît dans des expressions telles que *mise en place*. Enfin, nous enlevons le nom *plus*, car il est le plus souvent utilisé en tant qu'adverbe. Il nous reste donc 87 noms, que nous joignons en annexe[1]. Nous retrouvons sans surprise des mots tels que *hypothèse* et *résultat*.

1 Annexe 3 : Liste des noms retenus

Partie 3

L'expérimentation

Chapitre 7 – Les relations à analyser

Jaccard

Nous avons décidé d'utiliser sur notre corpus la méthode « distributionnelle ». Cette méthode utilise des mesures statistiques, plus précisément des mesures de distances entre un mot et un groupe de mots, ou entre groupes de mots. Cette distance est traditionnellement mesurée en utilisant l'indice de Jaccard ou la similarité cosinus.

L'indice de Jaccard est le rapport entre la cardinalité (la taille) de l'intersection des ensembles considérés et la cardinalité de l'union des ensembles. La similarité cosinus, beaucoup utilisée en « text mining », utilise le rapport entre le produit scalaire des ensembles et la taille des ensembles. Nous testerons les deux mesures.

Nous choisissons la mesure du Jaccard. Nous prenons donc deux à deux les noms de la liste précédemment créée. Nous observons pour les deux mots leurs contextes lexico-syntaxique et quelle partie de ce contexte ils partagent. Nous appliquons ensuite la formule du jaccard : jaccard = contexteCommun / (contexteMot1 + contexteMot2 - contexteCommun).

Prenons un exemple. Dans un petit texte, nous avons les associations *faire un plan*, *regarder un plan*, *un plan de travail* et les associations *demander une étude*, *faire une étude*. Le contexte du mot *plan* contient trois mots : *faire*, *regarder* et *travail*. Le contexte du mot *étude* contient les mots *demander* et *faire*. Les deux ont en commun le mot *faire*. Jaccard = 1 / (3 + 2 -1) = 1 / 4 = 0,25.

Cliques

Nous avons donc obtenu, pour chaque couple de mots de la liste calculée, une mesure de similarité.

Nous utilisons ces mesures pour créer des cliques de mots sémantiquement apparentés. Une clique est un ensemble de mots dont tous les membres sont reliés à chacun des autres par une relation donnée, ou un ensemble de relations. La relation étant présentement la proximité donnée par la mesure du jaccard.

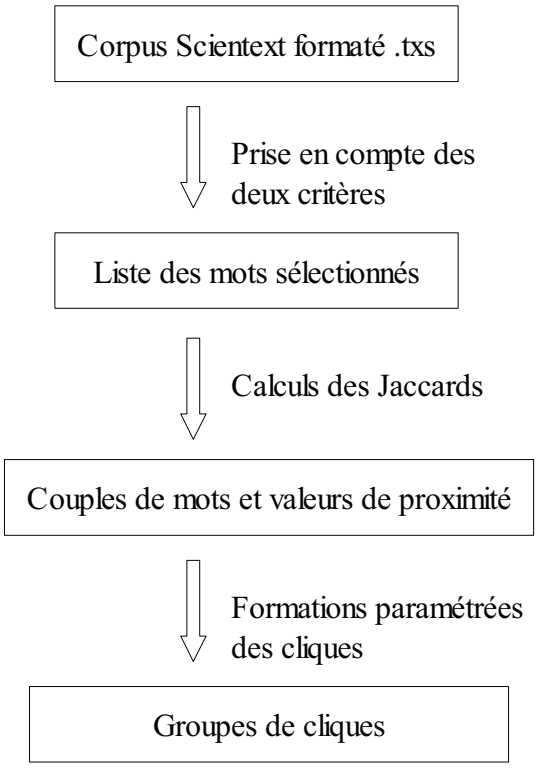


Figure 7 : Les étapes de l'expérimentation

Paramètres

Cette étude a pour but de confirmer ou d'infirmer nos hypothèses, c'est à dire de vérifier que le contexte lexicosyntaxique des mots peut aider à les rapprocher sémantiquement, que certaines relations syntaxiques sont plus pertinentes que d'autres et que la position des mots dans la structure du document est importante.

Pour étudier la première hypothèse, nous calculons des cliques de mots en utilisant toutes les relations extraites du corpus. Pour vérifier les deux autres hypothèses, nous calculons des cliques en rajoutant des critères.

Notre deuxième groupe de cliques sera calculé en ne prenant en compte que les relations sujet. Nous calculerons un troisième groupe de cliques en ne gardant que les relations objet. Nous obtiendrons aussi un quatrième groupe de cliques à partir des relations n'étant présentes que dans les résumés et les titres contenus dans les fichiers du corpus. Enfin, nous calculerons un dernier groupe de cliques ne tenant compte que des relations présentes dans au moins 5 des 9 disciplines du corpus. Cette dernière ne montrera rien quant à nos trois hypothèses, mais elle pourra nous indiquer si la transdisciplinarité des relations est nécessaire.

Chaque groupe de cliques teste donc un paramètre unique, nous pensons qu'ainsi les résultats seront plus visibles. Nous aurions pu n'utiliser que des pondérations. Par exemple faire varier un poids sur les relations sujet, et voir l'influence que cela a sur les résultats.

Ces groupes de cliques seront évalués deux fois. La première fois, les cliques seront formées en tenant compte des 100 couples de mots de Jaccard les plus forts. Par exemple, pour le groupe de cliques tenant compte de toutes les relations, nous trions les couples de deux mots par Jaccard décroissant, et nous prenons les 100 premiers. Nous n'imposons donc pas de seuils de Jaccard. Les résultats sont observables en annexe[1].

Nous choisissons d'attribuer aléatoirement un numéro sur les groupes de cliques. Le groupe n°1 sera celui obtenu avec les relations objet. Le groupe n°2 sera celui obtenu avec les relations sujet. Le n°3 sera celui obtenu en ne prenant en compte que le résumé et les titres. Le n°4 sera celui obtenu avec toutes les relations et enfin le n°5 sera celui qui respecte notre règle de transdisciplinarité. Nous affecterons les mêmes numéros pour les groupes de cliques calculés la seconde fois.

La seconde fois, nous mettons un seuil à la valeur du Jaccard. Ce seuil est égal au minimum entre 0,3 et la moitié du plus fort Jaccard de la catégorie. Nous pourrons ainsi voir si ne garder que des Jaccard « forts » rend les résultats meilleurs. L'inconvénient est que le nombre de cliques obtenues est restreint, voire vraiment trop faible dans le cas du groupe 3. Voici les résultats obtenus, à opposer aux résultats donnés en annexe 4.

GROUPE 1 : Cliques obtenues en ne gardant que les relations objet

GROUPE 1 Analyse Etude Recherche Travail
Cas Phénomène Problème
Modèle Situation Système
Niveau Nombre Valeur
Ensemble Objet Relation
Ensemble Objet Terme
Ensemble Relation Structure
Forme Relation Structure

1 Annexe 4 : Les groupes de cliques venant des 100 plus forts Jaccard

GROUPE 2 : Cliques obtenues en ne gardant que les relations sujet

GROUPE 2 Approche Etude Hypothèse
Activité Méthode Situation
Approche Modèle Système
Analyse Auteur Etude Travail
Analyse Etude Expérience Recherche
Elément Objet Relation
Elément Forme Relation
Elément Objet Terme
Elément Forme Structure Terme
Analyse Etude Résultat Travail
Analyse Approche Etude Modèle Recherche Travail

GROUPE 3 : Cliques obtenues en ne gardant que les relations présentes dans le résumé et les titres

GROUPE 3 Etude Expérience Recherche
Analyse Etude Recherche Travail

GROUPE 4 : Cliques obtenues en gardant toutes les relations

GROUPE 4 Elément Information Structure
Activité Processus Système
Analyse Approche Etude Modèle
Analyse Approche Etude Travail
Analyse Etude Recherche Travail
Elément Forme Objet Structure Terme Unité

GROUPE 5 : Cliques obtenues en ne gardant que les relations communes à au moins 4 des 9 disciplines.

GROUPE 5 Approche Méthode Modèle
Elément Facteur Point
Etude Recherche Travail
Expérience Recherche Travail
Caractéristique Elément Facteur
Expérience Méthode Test
Caractéristique Elément Structure

Figure 8 : Les cliques obtenues avec une restriction sur le Jaccard

Chapitre 8 – Les résultats

La première fois que nous avons créé des cliques, nous en avons obtenu un grand nombre, trop grand nombre qui contenait le mot *type*. Après réflexion, nous nous sommes dit que nous aurions du l'enlever lorsque nous avions restreint la liste des noms. En effet, *type* est un nom « approximatisant ». Il apparaît dans des phrases comme « *il utilise un type de bécher* ». Dans cette phrase, le sens se situe autour de la relation utilise-bécher, et non autour du mot *type*. C'est un mot sémantiquement vide que nous aurions dû enlever pendant les prétraitements. Nous ôtons donc ce mot de la liste et recréons les cliques.

Evaluation

Afin d'évaluer nos résultats, nous les soumettrons à des évaluateurs humains et appliquerons la règle de l'accord inter-annotateur. Nous soumettrons les différents groupes de cliques aux annotateurs, en nous gardant bien de leur indiquer les paramètres utilisés pour les obtenir. La « validité » de chaque groupe est évaluée par ces annotateurs, et nous pourrons en déduire un degré de pertinence de chacun de nos paramètres, répondant ainsi à notre problématique. Nous avons quatre évaluateurs, plus moi même. Ces quatre évaluateurs ont une très bonne connaissance du domaine, la plupart étant chercheur en linguistique. Ils sont donc qualifiés pour évaluer les résultats.

Ces évaluateurs affecteront à chaque clique une valeur de 1 à 5, 5 voulant dire que les termes de la clique sont proches. Comment définir cette proximité? Nous pensions originellement demander aux évaluateurs de considérer la quasi-synonymie des termes de la clique. Mais c'est une relation un peu forte, que l'analyse distributionnelle ne produit pas si souvent. Nous leur avons donc demandé de juger la proximité des mots en se demandant notamment s'ils ont un hypéronyme commun pas trop éloigné. Cependant, il est vrai que cette notation, d'après le retour des évaluateurs, se fait un peu à « l'intime conviction », au « feeling ». Il est très difficile d'évaluer la proximité de certains mots hors contexte.

Il serait intéressant d'étudier dans un autre projet, car nous manquons de temps, les manières d'évaluer ces cliques de manière plus précise, plus formelle. Il faudrait définir un encadrement plus strict à la notation, en veillant bien sûr à ne pas influencer la décision des évaluateurs.

Interprétations

Nous présentons ci-dessous les résultats des groupes de grandes cliques, formées en tenant compte des 100 couples de mots de Jaccard le plus forts.

	Evaluateur 1	Evaluateur 2	Evaluateur 3	Evaluateur 4	Moyenne
Groupe 1	2,25	3,08	2,21	1,96	2,38
Groupe 2	2,71	3	2,24	2,62	2,64
Groupe 3	1,94	3,09	1,5	2,53	2,27
Groupe 4	2,75	3,25	1,88	2,38	2,57
Groupe 5	2,57	3,43	1,57	2,43	2,5

Meilleur(s) score(s)
Score(s) moyen(s)
Plus mauvais score(s)

Groupe 1 : Cliques obtenues en ne gardant que les relations objet
Groupe 2 : Cliques obtenues en ne gardant que les relations sujet
Groupe 3 : Cliques obtenues en ne gardant que les relations présentes dans le résumé et les titres
Groupe 4 : Cliques obtenues en gardant toutes les relations
Groupe 5 : Cliques obtenues en ne gardant que les relations communes à au moins 4 des 9 relations

Figure 9 : Les résultats pour les groupes de grandes cliques

Nous pouvons être étonné de voir des valeurs osciller entre 2,27 et 2,64/5 en moyenne, mais elles sont en fait relativement satisfaisantes. Nous avons évoqué précédemment la nécessité de formaliser un système de notation, il faudra alors tenir compte des notes à attribuer, en plus de la façon de les attribuer.

Ces notes dans la moyenne ont tendance à valider notre première hypothèse, qui dit que l'on peut rapprocher sémantiquement des mots qui ont le même contexte lexico-syntaxique. Mais plus important que les notes elles mêmes, il est intéressant de les comparer entre elles, de regarder si certains groupes obtiennent des notes bien supérieures à d'autres.

Notre deuxième hypothèse consistait à dire que certaines relations sont plus pertinentes que d'autres pour évaluer la proximité sémantique entre plusieurs termes. Nous avions pris pour exemple les relations sujet et objet, c'est pourquoi nous nous sommes servi de ces deux là pour calculer les groupes 1 et 2. Nous voyons immédiatement que le groupe 1, celui des relations objet, obtient un des plus mauvais scores, alors que le groupe 2, celui des relations sujet, obtient le meilleur score.

Nous pouvons donc observer que la relation sujet est plus pertinente que la relation objet, ce qui est l'inverse de ce que trouvait Cécile Fabre [Fabre, 2010].

Trois des quatre annotateurs mettent la meilleure note au groupe des relations sujet, et ce malgré la difficulté de décerner les notes. Nous pouvons donc considérer cette déduction comme pertinente, même si un des évaluateurs a donné sa moins bonne note à ce groupe.

Comparons les résultats du groupe 2 avec celui du groupe 4, qui avait été créé à partir de toutes les relations, et qui sert donc d'échantillon test. Le groupe 2 obtient de meilleurs résultats. Cela montre encore une fois que la relation sujet serait une des plus pertinentes.

Regardons le groupe numéro 5, correspondant aux couples obtenus avec seulement des relations respectant notre critère de transdisciplinarité, des relations présentes dans au moins 4 des 9 disciplines. Il obtient un résultat assez proche de celui du groupe 4, mais moins bon. Cette observation reste mitigée, car deux des quatre annotateurs, soit la moitié, donne une moins bonne note au groupe 5, les deux autres donnant une moins bonne note au groupe 4. On peut en déduire qu'il serait inutile de ne conserver que les relations transdisciplinaires, car elles semblent détériorer le résultat. Rappelons quand même que les noms choisis restent eux transdisciplinaires.

Enfin, regardons le groupe numéro 3, celui correspondant aux couples obtenus avec les relations présentes dans les résumés et les titres. Il obtient le plus faible résultat. Deux annotateurs sur quatre le considère comme étant le moins bon groupe, un troisième utilisateur le classe parmi les moins bons. Cela infirmerait notre troisième hypothèse, qui disait au contraire que l'on devrait obtenir un meilleur résultat. Ceci est sans doute dû au fait qu'il n'y ait pas de relations sujet et objet dans les titres. En effet, nous venons d'observer que les relations sujet sont les plus pertinentes. Il est donc logique que si on les enlève, ce qui est le cas dans les titres qui n'en n'ont pas ou très peu, le résultat est faible.

Malgré la difficulté de noter les groupes de cliques, les annotateurs ont, sans se concerter, jugé le groupe des relations sujet comme semblant être le meilleur et le groupe des relations extraites des titres et résumé comme semblant être le moins bon.

Essayons de vérifier ces déductions avec les cliques obtenues en seuillant le jaccard des couples considérés.

	Evaluateur 1	Evaluateur 2	Evaluateur 3	Moyenne
Groupe 1	2,25	3,25	2,88	2,79
Groupe 2	2,36	3,45	3,27	3,03
Groupe 3	4,5	5	5	4,83
Groupe 4	3,33	3,5	3,67	3,5
Groupe 5	3,14	4,43	4,14	3,9

Meilleur(s) score(s)
Score(s) moyen(s)
Plus mauvais score(s)

Groupe 1 : Cliques obtenues en ne gardant que les relations objet
Groupe 2 : Cliques obtenues en ne gardant que les relations sujet
Groupe 3 : Cliques obtenues en ne gardant que les relations présentes dans le résumé et les titres
Groupe 4 : Cliques obtenues en gardant toutes les relations
Groupe 5 : Cliques obtenues en ne gardant que les relations communes à au moins 4 des 9 relations

Figure 10 : Les résultats pour les groupes de petites cliques

Avant d'utiliser les résultats de la figure 10, il convient de préciser que ces résultats proviennent de groupes de cliques bien plus petits que pour l'expérience précédente. Le groupe 3, notamment, est inutilisable car il n'est composé que de deux cliques, ce qui est trop peu. Nous ne tiendrons donc pas compte du groupe 3.

La première observation que l'on peut faire est que l'on obtient pour tous les groupes une note moyenne supérieure à celle obtenue précédemment. Il semble donc qu'il soit plus intéressant de seuiller le jaccard, en choisissant une valeur minimum. Toute la difficulté sera alors de choisir un minimum assez grand pour obtenir des valeurs finales hautes, mais assez petites pour entraîner la création d'un nombre suffisant de cliques, pour éviter le cas du groupe 3, présentement. Il ne faut également pas oublier qu'avec un seuil sur le jaccard, nous conservons moins de relations. Les cliques sont donc moins nombreuses mais surtout moins grandes, cela entraîne de meilleures notes, par effet de bord. En effet, plus une clique est grande, plus il y a de chances qu'un ou plusieurs des termes qui la composent ne soit pas très cohérent avec les autres.

Ces résultats confirment que la relation objet ne serait pas très utile, les trois annotateurs la jugeant comme la moins pertinente, mais ils montrent aussi que la relation sujet, même si elle semble là encore plus pertinente que la relation objet, serait moins efficace que ce que l'on vient de voir.

En effet, plus aucun annotateur ne donne sa meilleure note au groupe 2.

Il nous faudrait demander l'avis de plus d'évaluateurs pour voir si les deux valeurs se rapprocheraient.

Nous accorderons un peu plus de foi à la première expérience car la valeur a été calculée sur un bien plus grand nombre de cliques.

Nous voyons aussi que si l'on seuille les jaccard, pour ne garder que les très forts, l'utilisation de seulement les relations transdisciplinaires donnent de meilleurs résultats. Deux des trois annotateurs affectent à ce groupe la meilleure note, le troisième lui donnant quand même une des meilleures notes.

Le bilan de ces deux expérimentations laisse penser que contrairement à ce que nous pensions, la position des mots dans la structure du document n'aurait pas d'incidence sur le calcul de proximités entre différents noms voire qu'elle aurait une influence néfaste. Le caractère transdisciplinaire des relations mises en jeu n'affecterait pas non plus ce calcul.

Par contre, nous vérifions que certaines relations, ici la relation sujet, serait plus pertinentes que d'autres, ici la relation objet.

Et bien sûr, nous voyons qu'utiliser le contexte lexico-syntaxique des mots permet de les rapprocher sémantiquement. On pourra alors comparer les résultats obtenus avec les résultats obtenus sans tenir compte de la syntaxe, pour voir s'ils sont meilleurs.

Conclusion

Nous regrettons de ne pas avoir eu plus d'évaluateurs, pour pouvoir accorder encore plus de crédits à nos résultats, mais nous avons quand même des résultats.

Nous avons pu confirmer certaines hypothèses, en infirmer d'autres. Dans les deux cas, les résultats sont utiles mais soulèvent de nouvelles questions. Pour le cas des relations plus pertinentes que d'autres, nous en avons testé ici quelques unes, il faudrait toutes les tester. Nous avions pris comme hypothèse forte le fait que certaines relations étaient inutiles, de point de vue sémantique, notamment la relation déterminant. Une étude pourrait les reprendre en compte pour le vérifier, ce qui entraînera une augmentation dans les temps d'exécution et la taille des différents fichiers de sortie.

Nous avons pu voir que contrairement à ce que l'on pensait, la position des mots dans la structure des documents n'aurait pas d'incidence sur le fait de les rapprocher sémantiquement. Il s'avèrerait donc inutile d'effectuer un lourd travail d'annotation, de formalisation de documents pour garder les informations de structure, en tout cas pour les études sur le lien sémantique entre les mots.

Bibliographie

Aït-Mokhtar, S., Chanod, J-P. & Roux, C. (2001). A multi-input dependency parser. *Dans Proceedings of the the Seventh International Workshop on Parsing Technologies*. Beijing : Tsinghua University Press.

Bourigault, D. (2002). UPERY : Un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002)* (pp. 75-84). Le Chesnay, France : INRIA.

Bourigault, D. & Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire* (pp 131-151). Toulouse : ERSS.

Fabre, C. (2010). *Affinités syntaxiques et sémantiques entre les mots - apports mutuels de la linguistique et du traitement automatique des langue.* Université Toulouse 2 – Le mirail et CNRS.

Galy, E. & Bourigault, D. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. *In G. Williams, éd., Texte et corpus: Actes des 4e journées de linguistique de corpus (JLC)* (pp 163-174). Lorient.

Grefenstette, G. (1996). *Evaluation techniques for automatic semantic extraction : Comparing syntactic and window based approaches* (pp 143-153) Cambridge, Massachusetts : MIT Press.

Harris, Z., S. (1968). *Mathematical structures of language.* New-York, NY: John Wiley & Sons.

Hagège, C. & Roux, C. (2003). Entre syntaxe et sémantique : Normalisation de la sortie de l'analyse syntaxique en vue de l'amélioration de l'extraction d'information à partir de textes. *Actes de TALN, TALN 2003*. Batz-sur-Mer, France.

Lin, D. & Pantel, P. (2001). *Induction of semantic classes from natural language text.* New York, NY.

Phillips, M. (1985). *Aspects of text structure: An investigation of the lexical organization of text.* Amsterdam, Pays-Bas: Elsevier.

Tutin, A. (2007). Traitement sémantique par analyse distributionnelle des noms transdisciplinaires des écrits scientifiques. *Actes de TALN 2007* (pp 283-292). Toulouse.

Tutin, A. (2007). Autour du lexique et de la phraséologie des écrits scientifiques. *Revue Française de Linguistique Appliquée. Volume XII-2* (pp 5-14).

Table des annexes

Annexe 1
Un exemple de fichier de Scientext

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="fr-FR">
   <?oxygen RNGSchema="file:teilite.rnc" type="compact"?>
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>corpus scientext</title>
         </titleStmt>
         <publicationStmt>
            <publisher>LIDILEM -Projet Scientext -Université Grenoble 3 -Stendhal -F.
Grossmann (francis.grossmann@u-grenoble3.fr)
               -A. Tutin (agnes.tutin@u-grenoble3.fr)</publisher>
            <publisher>www.u-grenoble3.fr/lidilem/scientext</publisher>
            <address>
               <addrLine>Université Stendhal Grenoble III -UFR des Sciences du Langage -
BP 25 -38040 Grenoble cedex 9
                  Tél. : +33 (0) 4 76 82 43 74 -Fax : +33 (0) 4 76 82 43 95</addrLine>
            </address>
            <date>février 2009</date>
            <idno>texte_article_scientext_87</idno>
            <distributor>LIDILEM</distributor>
            <availability>
               <p>Disponible gratuitement sous licence Creative Commons :
http://creativecommons.org/licenses/by/2.0/</p>
               <p>Cette licence implique la libre reproduction et diffusion de l'œuvre au
public, notamment par téléchargement, mais est soumise aux conditions suivantes :</p>
               <p>-        Respect du droit à la Paternité : l'Utilisateur est tenu de citer le
nom de l'auteur original de la manière indiquée par l'Auteur de l'oeuvre (mais pas d'une
manière qui suggérerait que l'Auteur soutient ou approuve l'utilisation de l'œuvre par
l'Utilisateur) ;</p>
               <p>-        Pas d'Utilisation Commerciale : l'Utilisateur ne peut exploiter ou
diffuser l'œuvre à des fins commerciales ;</p>
               <p>-        Pas de Modification : l'Utilisateur ne peut modifier l'œuvre
originale sans l'accord express de l'Auteur, quelle que soit la nature et l'importance de
cette modification ;</p>
               <p>-        Partage des Conditions à l'Identique : L'Utilisateur ne peut diffuser
l'œuvre que sous une licence identique ; et aux mêmes conditions.</p>
            </availability>
         </publicationStmt>
         <sourceDesc>
           [...]
            <keywords scheme="genre">
               <list><item>article</item></list>
            </keywords>
            <keywords scheme="discipline"><list><item>linguistique</item></list>
            </keywords>
[...]
```

```
  </teiHeader>
    <text>
    <front>
      <head>Figures et référence plurielle,<lb/>en corpus journalistique</head>
      <docAuthor>
        <name>Michelle Lecolle<note place="foot" n="*"> ERSS (UMR 5610, CNRS /
Université
            Toulouse II) et Université Toulouse II.</note></name>
      </docAuthor>
      <div type="abstract">
        <p>Nous étudions ici des énoncés figuraux basés sur l'expression de références
          plurielles (syntagmes nominaux définis pluriel, noms collectifs), dans un
corpus
          journalistique portant sur l'actualité. Est abordée tout d'abord la description
          d'une figure que nous nommons énallage, et qui se présente comme un jeu
entre
          appréhension collective et distributive d'un ensemble. Puis celle de
          l'hyperbole, qui se manifeste ici dans l'amplification du nombre. Enfin nous
          examinons le trope synecdochique substituant le tout à la partie (ce qui
          correspond ici à un rapport ensemble/sous-ensemble). Nous prenons en
compte dans
          la description le cadre énonciatif de la communication journalistique et
          considérons l'usage argumentatif que l'énonciateur peut faire des figures
          citées.</p>
        </div>
      <div type="abstract" xml:lang="en">
        <p>In this paper I describe three figures taken from a journalistic corpus. These
          figures are based on the expression of plural reference (definite plural noun
          phrases and collective nouns). The first figure, which I call énallage, involves
          in this context a grammatical interplay between distributive and collective
          reference. The second figure studied is hyperbole, here amplification in
number.
          I then examine a synecdoche which substitutes the whole for the part (here
          corresponding to a set/subset relation). The journalistic context of
          communication forms part of the description, and I also take into account the
          argumentative use that speakers can make of these figures.</p>
        </div>
    </front>
    <body>
      <div type="introduction" n="1">
        <head>Introduction</head>
        <p>La tradition rhétorique nous a transmis (par l'intermédiaire en particulier de
          Dumarsais et de Fontanier) un répertoire raisonné de procédés stylistiques,
          nommés figures et tropes. Les procédés cités sont généralement considérés
comme
          devant donner « plus d'ornement », « plus de vivacité » au discours. Mais ils
          sont décrits hors de tout contexte ou, si le contexte est pris en compte, les
          exemples donnés sont le plus souvent tirés de textes littéraires, ou même de
          textes latins (Dumarsais). C'est cette absence d'ancrage actuel qui peut
amener
          à considérer leur exploitation de nos jours comme « un anachronisme
            stérile »<ref target="#_ftn1">[1]</ref>.
```

[...]

Annexe 2
Exemple de contenu d'un fichier TXS


```
<?xml version="1.0" encoding="utf-8"?>
<txs file="article_100_ling_LIDIL_Niederberger__Berthoud_Papandropoulou.txs">
<chunkList>
<chunk>
<idno>100</idno>
<genre>article</genre>
<discipline>linguistique</discipline>
<text >
<front >
<head >
        <s id="s1">
                <tokens>
                        <t id="t1" i="0" l="utilisation" f="Utilisation"
p="NOUN" c="PARSN DESN STARTBIS MAJ SFPAR SFDE CLOSED FEM SG P3
NOUN START LAST FIRST"/>
                        <t id="t2" i="2" l="de" f="des" p="PREP" c="SFDE
FORM MASC FEM PL DEF PREP DET FIRST"/>
                        <t id="t3" i="4" l="pronom" f="pronoms" p="NOUN"
c="CLOSED MASC PL P3 NOUN LAST FIRST"/>
                        <t id="t4" i="6" l="personnel" f="personnels" p="ADJ"
c="MASC PL ADJ LAST FIRST"/>
                        <t id="t5" i="8" l="en" f="en" p="PREP" c="SFEN
FORM PREPANNEE PREP FIRST"/>
                        <t id="t6" i="10" l="français" f="français" p="NOUN"
c="MASC PL SG P3 NAT NOUN LAST FIRST"/>
                        <t id="t7" i="12" l="écrit" f="écrit" p="ADJ"
c="SURSN QUEP ENSN IMPERSO SFSUR SFEN MASC SG ADJ LAST FIRST"/>
                        <t id="t8" i="14" l="par" f="par" p="PREP"
c="PREPINF SFPAR FORM PREP FIRST"/>
                        <t id="t9" i="16" l="un" f="des" p="DET" c="FORM
CLOSED MASC FEM PL INDEF DET FIRST"/>
                        <t id="t10" i="18" l="enfant" f="enfants" p="NOUN"
c="CLOSED MASC FEM PL P3 NOUN LAST"/>
                        <t id="t11" i="20" l="sourd" f="sourds" p="ADJ"
c="ASN SFA MASC PL ADJ LAST FIRST"/>
                        <t id="t12" i="22" l="bilingue" f="bilingues" p="ADJ"
c="MASC FEM PL ADJ CR LAST FIRST"/>
                        <t id="t13" i="24" l=":" f=":" p="PUNCT" c="FORM
STRONGBREAK TOUTMAJ1 PUNCT"/>
                        <t id="t14" i="26" l="un" f="un" p="DET"
c="STARTBIS CLOSED MASC SG TOUTMAJ1 NUMROI INDEF LETTRES DET
FIRST"/>
                        <t id="t15" i="28" l="parcours" f="parcours"
p="NOUN" c="PARSN DESN SFPAR SFDE CLOSED MASC PL SG P3 NOUN LAST"/
>
                        <t id="t16" i="30" l="spécifique" f="spécifique"
p="ADJ" c="DESN SFDE MASC FEM SG ADJ LAST FIRST"/>
```

<t id="t17" i="32" l="de" f="d'" p="PREP" c="PREPINF SFDE FORM PREP DIR FIRST"/>
<t id="t18" i="34" l="apprentissage" f="apprentissage" p="NOUN" c="PARSN DESN SFPAR SFDE MASC SG P3 NOUN LAST FIRST"/>
<t id="t19" i="36" l="?" f="?" p="SENT" c="TOUTMAJ1 SENT END LAST"/>
</tokens>
<sentence>Utilisation(0) des(2) pronoms(4) personnels(6) en(8) français(10) écrit(12) par(14) des(16) enfants(18) sourds(20) bilingues(22) :(24) un(26) parcours(28) spécifique(30) d'(32) apprentissage(34) ?(36) </sentence>
<dependances>
<g r="ADJMOD" s="30" c="34" />
<g r="DE_NMOD" s="0" c="4" />
<g r="PAR_NMOD" s="10" c="18" />
<g r="EN_NMOD" s="4" c="10" />
<g r="ADJ_NMOD" s="28" c="30" />
<g r="ADJ_NMOD" s="4" c="6" />
<g r="ADJ_NMOD" s="10" c="12" />
<g r="ADJ_NMOD" s="18" c="20" />
<g r="ADJ_NMOD" s="18" c="22" />
</dependances>
</s>

Annexe 3
Liste des noms retenus

activité
an
analyse
année
approche
augmentation
auteur
base
capacité
caractéristique
cas
chapitre
choix
compte
condition
contrainte
contrôle
critère
différence
distance
domaine
donnée
développement
enfant
ensemble
erreur
expression
expérience
facteur
fait
façon
figure
fonction
forme
fréquence
groupe
hypothèse
information
interaction
intérêt
manière
mesure
mode
modèle
méthode
niveau
nombre
objectif
objet
observation
ordre
partie
phase
phénomène
point
position
principe
problème
processus
production
rapport
recherche
relation
référence
réponse
résultat
rôle
situation
structure
sujet
système
tableau
temps
terme
test
traitement
travail
unité
utilisation
valeur
variable
élément
état
étude
évidence
évolution

Annexe 4
Les groupes de cliques venant des 100 plus forts Jaccards

GROUPE 1 : Cliques obtenues en ne gardant que les relations objet

GROUPE 1
Analyse Approche Recherche
Cas Phénomène Problème
Analyse Etude Recherche Travail
Activité Processus Travail
Relation Résultat Valeur
Critère Modèle Principe
Critère Méthode Modèle
Méthode Modèle Système
Modèle Situation Système
Caractéristique Cas Point
Activité Niveau Nombre
Ensemble Objet Partie Unité
Phénomène Processus Relation
Phénomène Relation Situation
Elément Nombre Valeur
Ensemble Structure Système
Objet Relation Situation
Niveau Relation Situation
Ensemble Niveau Nombre Structure Valeur
Ensemble Niveau Relation Structure Valeur
Ensemble Niveau Relation Structure Unité
Ensemble Forme Objet Relation Structure
Ensemble Forme Objet Structure Terme
Ensemble Objet Relation Structure Unité

GROUPE 2 : Cliques obtenues en ne gardant que les relations sujet

GROUPE 2
Elément Forme Unité
Analyse Auteur Etude Travail
Approche Etude Hypothèse
Etude Hypothèse Résultat
Elément Objet Relation Terme
Elément Forme Relation Terme
Elément Forme Structure Terme
Elément Forme Nombre Structure
Forme Modèle Système
Analyse Approche Elément
Approche Etude Modèle Système
Activité Méthode Situation
Activité Production Situation
Activité Production Traitement
Activité Méthode Traitement
Activité Expérience Méthode
Analyse Etude Recherche Résultat Travail
Analyse Etude Expérience Recherche Travail
Analyse Ensemble Expérience Travail
Analyse Approche Ensemble Travail
Analyse Approche Etude Modèle Recherche Travail

GROUPE 3 : Cliques obtenues en ne gardant que les relations présentes dans le résumé et les titres

GROUPE 3 Etude Hypothèse Modèle Résultat
Analyse Etude Observation Recherche
Etude Méthode Modèle Résultat
Forme Mesure Modèle
Mesure Modèle Recherche
Relation Situation Structure
Analyse Approche Etude Modèle
Analyse Approche Modèle Structure
Développement Modèle Résultat
Fonction Forme Modèle
Fonction Modèle Objet
Activité Modèle Objet Situation
Niveau Système Traitement
Modèle Niveau Résultat
Forme Modèle Niveau Système
Activité Forme Modèle Situation Structure Système
Analyse Structure Système Traitement
Analyse Modèle Structure Système
Modèle Recherche Situation Structure
Etude Expérience Modèle Recherche Situation
Analyse Etude Recherche Traitement
Analyse Recherche Structure Traitement
Analyse Etude Expérience Modèle Résultat
Analyse Etude Expérience Résultat Travail
Analyse Etude Expérience Recherche Travail
Analyse Etude Expérience Modèle Recherche
Donnée Etude Résultat Travail
Donnée Etude Recherche Travail
Donnée Etude Modèle Recherche
Donnée Modèle Recherche Structure
Donnée Etude Modèle Résultat
Donnée Modèle Résultat Structure
Analyse Modèle Résultat Structure
Analyse Modèle Recherche Structure

GROUPE 4 : Cliques obtenues en gardant toutes les relations

GROUPE 4 Expression Structure Terme
Etude Expérience Travail
Critère Elément Facteur
Critère Hypothèse Principe
Caractéristique Elément Structure
Caractéristique Fonction Structure
Fonction Relation Structure
Fonction Structure Valeur
Elément Information Structure Unité
Nombre Résultat Valeur
Activité Processus Système
Activité Processus Travail
Forme Relation Structure
Analyse Résultat Travail
Elément Forme Structure Valeur
Activité Objet Situation Système
Objet Situation Système
Modèle Situation Système
Analyse Approche Etude Recherche Travail
Analyse Approche Etude Travail
Analyse Approche Etude Modèle
Elément Ensemble Structure Terme
Forme Modèle Structure Système
Forme Objet Structure Système
Elément Forme Objet Structure Terme Unité

GROUPE 5 : Cliques obtenues en ne gardant que les relations communes à au moins 4 des 9 disciplines.

GROUPE 5 Analyse Etude Travail
Activité Processus Traitement
Activité Processus Système
Caractéristique Elément Facteur Point
Caractéristique Elément Point Problème
Caractéristique Expérience Point
Etude Expérience Recherche Travail
Approche Méthode Test
Expérience Méthode Test
Approche Méthode Mode
Approche Méthode Modèle
Méthode Modèle Structure Système
Caractéristique Elément Fonction Structure
Expérience Groupe Situation
Expérience Groupe Méthode
Caractéristique Elément Objet Structure
Elément Forme Objet Structure
Forme Structure Système
Forme Objet Situation Structure
Activité Expérience Objet Situation Structure
Activité Expérience Méthode Structure Système
Activité Expérience Méthode Objet Structure
Activité Caractéristique Expérience Objet Structure

Annexe 5
Les groupes de cliques utilisant des Jaccards seuillés

GROUPE 1 : Cliques obtenues en ne gardant que les relations objet

GROUPE 1 Analyse Etude Recherche Travail
Cas Phénomène Problème
Modèle Situation Système
Niveau Nombre Valeur
Ensemble Objet Relation
Ensemble Objet Terme
Ensemble Relation Structure
Forme Relation Structure

GROUPE 2 : Cliques obtenues en ne gardant que les relations sujet

GROUPE 2 Approche Etude Hypothèse
Activité Méthode Situation
Approche Modèle Système
Analyse Auteur Etude Travail
Analyse Etude Expérience Recherche
Elément Objet Relation
Elément Forme Relation
Elément Objet Terme
Elément Forme Structure Terme
Analyse Etude Résultat Travail
Analyse Approche Etude Modèle Recherche Travail

GROUPE 3 : Cliques obtenues en ne gardant que les relations présentes dans le résumé et les titres

GROUPE 3 Etude Expérience Recherche
Analyse Etude Recherche Travail

GROUPE 4 : Cliques obtenues en gardant toutes les relations

GROUPE 4 Elément Information Structure
Activité Processus Système
Analyse Approche Etude Modèle
Analyse Approche Etude Travail
Analyse Etude Recherche Travail
Elément Forme Objet Structure Terme Unité

GROUPE 5 : Cliques obtenues en ne gardant que les relations communes à au moins 4 des 9 disciplines.

GROUPE 5 Approche Méthode Modèle
Elément Facteur Point
Etude Recherche Travail
Expérience Recherche Travail
Caractéristique Elément Facteur
Expérience Méthode Test
Caractéristique Elément Structure

Table des illustrations

Sigles et abréviations utilisés

Sigle : TAL Traitement Automatique des Langues, TEI Text Encoding Initiative, TXS format xml étiqueté, XIP Xerox Incremental Parser, XML eXtensible Markup Language

Table des matières

MOTS-CLÉS : Traitements automatiques des langues, analyse syntaxique, classes sémantique, approche distributionnelle.

RÉSUMÉ

Nous proposons une étude exploratoire autour de l'approche distributionnelle (Harris, 1968). Dans ce cadre général, nous étudierons plus en détail l'influence de différents facteurs sur le rapprochement sémantique entre mots. Notre étude sera accée sur deux niveaux. Nous souhaitons d'abord étudier l'influence des différents types de relations syntaxiques, leur pertinence. Par exemple, la relation sujet est peut être plus utile que la relation objet pour rapprocher sémantiquement des mots. Nous observerons aussi la position des mots dans la structure du document, ceci à deux niveaux. A la fois sur la structure global du document, pour voir si le texte présent dans le résumé est plus important que le texte présent dans le corps du texte, la conclusion, ou autre. Mais aussi, à l'intérieur d'une section, nous voulons savoir s'il est intéressant de s'intéresser plus aux titres qu'aux paragraphes qui les suivent.

KEYWORDS : Language processing, syntactic analysis, semantic classes, distributional approach.

ABSTRACT

We present an explorative study about the distributional approach (Harris, 1968). In this general framework, we will investigate the influence of various factors on the semantic reconciliation between words. Our study will focus on two levels. We will start by study the influence of different types of syntactic relations, their relevance. For example, subject relationship may be more useful than the object relationship to semantically gather words. We will also observe the position of words in the structure of the document, on two levels. On the global structure of the document, to see if sentences from abstract are more important than the sentences found in sections, conclusion, or anaywhere else. But also in a section, we want to know if it is interesting to pay more intention to titles rather than the following paragraphs.

UNIVERSITE Stendhal GRENOBLE 3

Déclaration anti-plagiat

Document à scanner après signature
et à joindre au mémoire électronique

DECLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : GALOP PRENOM : MICKAEL

DATE : 26/09/11