Le traitement automatique : un moteur pour l'évolution des dictionnaires de synonymes

Jean-Luc Manguin, Lonneke van der Plas, Jörg Tiedemann

Le traitement automatique : un moteur pour l'évolution des dictionnaires de synonymes

Jean-Luc Manguin (1)
jean-luc.manguin@unicaen.fr
Lonneke Van der Plas (2)
vdplas@let.rug.nl
Jörg Tiedemann (2)
tiedeman@let.rug.nl

(1) CRISCO, Université de CAEN (FR)
(2) Alfa-Informatica, Université de GRONINGEN (NL)

Mots-clés : lexicographie, extraction de synonymes, corpus multilingues.

Résumé : Nous présentons ici un aperçu des traitements automatiques liés aux dictionnaires de synonymes et devant servir à l'enrichissement lexicographique. Nous décrivons tout d'abord les méthodes endogènes qui se basent sur une modélisation en graphe de la relation de synonymie, et qui ont pour but secondaire de vérifier la qualité des relations. En seconde partie, nous abordons les méthodes exogènes qui extraient les relations synonymiques d'une analyse de corpus textuels monolingues ou multilingues. Nous insisterons plus particulièrement sur ce second type de corpus qui apporte une précision et un rappel supérieurs à la méthode monolingue, et dégage de bonnes perspectives lexicographiques.

Introduction : la lexicographie instrumentée

La préface du récent « Dictionnaire des combinaisons de mots » paru aux éditions Le Robert [Le Fur, 2006] nous révèle l'importance actuelle d'une lexicographie « à l'instrument »[1], y compris dans l'élaboration de produits destinés à l'édition papier. Dans le domaine des dictionnaires électroniques, la partie instrumentale devient parfois un but en soi, concomitant à l'élaboration de la ressource, comme dans les travaux qui visent à construire des ressources terminologiques.
Dans le cas des dictionnaires de synonymes, l'instrumentation telle que nous venons de la décrire aura aussi pour but la construction de la ressource, ou d'une partie complémentaire de celle-ci. Cependant, il importe de distinguer deux aspects dans cette construction, qui servent d'ailleurs à disposer les dictionnaires selon deux points de vue classiquement admis[2] : d'une part la recherche des substituts d'un mot (qui oriente la ressource constituée vers les dictionnaires cumulatifs), d'autre part la recherche des conditions de substitution (qui la place

[1] En extrapolant l'expression de B. Habert [Habert, 2005].
[2] Voir par exemple [Quemada, 1968], ou [Pruvost, 2007].

cette fois du côté des dictionnaires distinctifs). Il va de soi que le second aspect exige de recourir à un matériau externe à la ressource que l'on bâtit, qu'il soit un corpus de textes ou un dictionnaire complémentaire. Par contre, et c'est ce que nous allons développer ici, l'ajout de nouvelles relations synonymiques à un dictionnaire déjà existant peut non seulement faire appel à des apports extérieurs, mais aussi fonctionner de manière autarcique. Dans les deux cas, la théorie sous-jacente diffère sérieusement en raison de la nature des données à traiter, et du modèle qui leur est appliqué.

1. Les forces et les faiblesses d'un dictionnaire électronique

Il existe peu de dictionnaires électroniques des synonymes du français, tandis que l'anglais bénéficie des développements du projet WordNet, qui n'est pas à proprement parler un dictionnaire de synonymes, puisqu'il organise les mots en réseau au moyen de différents types de relations (entre autres synonymie, hyponymie, hyperonymie). Celui que nous avons utilisé dans ce travail est mentionné par Habert [op. cit.], et plus connu sous le nom de « Dictionnaire Electronique des Synonymes du CRISCO ». Sa particularité essentielle, voulue dès le départ par ses concepteurs, est d'être purement cumulatif, c'est-à-dire de ne donner pour chaque entrée qu'une liste de synonymes, sans faire de distinction d'emploi ni de regroupement de sens.

Cette démarche radicale s'explique par le fait que ce dictionnaire ne devait constituer qu'un support réel à une démarche de modélisation du sens formulée par B. Victorri [Victorri et Fuchs, 1996] ; la mise en ligne de cette ressource et le succès qui a suivi ont révélé par ailleurs la forte demande des internautes à l'égard d'une telle ressource, et ont déclenché des travaux complémentaires visant à doter la liste des synonymes d'informations supplémentaires, comme la barre de classement des synonymes.

Cela dit, le contenu initial du dictionnaire, issu de la fusion de fichiers contenant les liens synonymiques présents dans sept autres ressources[3], et complété par le travail d'harmonisation des lexicographes du CRISCO, reste néanmoins essentiellement tributaire des relations posées par les auteurs des dictionnaires compilés. Par exemple, la relation de synonymie entre *curieux* et *insolite*, mentionnée par le TLFi, n'est présente dans aucun des dictionnaires sources. Cet exemple simple montre, comme l'avait déjà signalé Kahlmann [Kahlmann, 1975], que d'une part le travail des lexicographes peut être entaché d'oublis, et que d'autre part, même la compilation de plusieurs ouvrages ne met pas à l'abri de telles lacunes.

D'un autre côté, la simplicité structurelle de ce dictionnaire de synonymes, dans lequel les mots sont simplement reliés par une relation symétrique de type booléen, permet de le faire entrer parmi les graphes. Placé ainsi dans une catégorie d'objets mathématiques dont le substrat théorique est déjà abondamment développé et toujours en cours d'enrichissement, il peut être soumis à des analyses et à des transformations bien connues, et dont les résultats pourront être examinés en tenant compte cette fois de la valeur sémantique de la relation présente dans le graphe. En d'autres termes, l'objectivité des méthodes mathématiques de la théorie des graphes fait contrepoids à la critique précédente concernant la subjectivité des liens présents dans la ressource.

[3] Dictionnaires de Guizot, Lafaye, Bailly, Bénac, du Chazaud, et renvois synonymiques du Grand Larousse et du Grand Robert ; pour plus de détails voir [Ploux, 1997].

2. Enrichissement endogène

Ce premier type d'enrichissement ne consiste bien entendu qu'à ajouter des liaisons dans le graphe de synonymie, puisqu'il se base sur l'exploitation de ce graphe par des méthodes diverses ; celles-ci ne peuvent pas faire apparaître de nouveaux nœuds dans la structure.
Le principe général de ces méthodes consiste à explorer le graphe de synonymie autour d'un nœud, en allant plus loin que les voisins directs de ce nœud de départ. Les premiers travaux de ce genre sont probablement ceux de Brodda et Karlgren [Brodda et Karlgren, 1969], qui ont proposé un modèle « thermodynamique » fondé sur une analogie entre le réseau synonymique et un réseau conducteur de chaleur avec des pertes de transmission. Les auteurs ont réalisé une version informatique d'un dictionnaire de synonymes suédois dans lequel, pour trouver les synonymes proches d'un mot, il suffisait de « chauffer » le mot en question, puis de laisser le système arriver à un état d'équilibre tout en maintenant constante la température du mot « chauffé » ; bien entendu, tous les changements de température au sein du réseau étaient calculés par l'ordinateur, et l'utilisateur n'avait plus qu'à consulter une liste des mots les plus « chauds », qui selon le modèle implémenté, s'avérait correspondre aux synonymes les plus proches du mot activé. L'intérêt du modèle était d'offrir la possibilité de chauffer plusieurs mots à la fois, par exemple dans le cas où l'un d'eux est polysémique, et de pouvoir faire varier la transmission de chaleur entre les nœuds du graphe ; mais la finalité n'était pas de modifier le réseau de départ.
Plus récemment, JL Manguin [Manguin, 2004] a proposé d'ajouter de nouvelles liaisons par « transitivité conditionnelle » ; cette méthode se base d'une part sur la transitivité (si B est synonyme de A, et C synonyme de B, alors C est synonyme de A), et d'autre part sur la similitude entre deux mots du graphe, mesurée par l'indice de Jaccard. En l'occurrence, cet indice de communauté est égal au nombre de synonymes communs aux deux mots, divisé par le nombre total de synonymes qu'ils possèdent à eux deux. La transitivité est dite « conditionnelle » si pour B synonyme de A et C synonyme de B, on peut ajouter une liaison entre A et C qu'à certaines conditions, notamment si la similitude entre A et C est strictement supérieure à 0,5 . Dans l'étude faite à partir du dictionnaire de Bailly, les liaisons ajoutées étaient très pertinentes, et si des liaisons proposées étaient aberrantes, cela révélait même des erreurs dans le dictionnaire d'origine. L'exemple mentionné précédemment de *curieux* et *insolite* peut être résolu de cette manière dans le DES du CRISCO.
Dans un but presque similaire, Bruno Gaume [Gaume, 2006] a exploité le graphe de synonymie issu du même dictionnaire de synonymes, pour trois catégories (verbe, nom et adjectif), par une méthode mathématique tenant compte de la totalité du graphe, afin de trouver quels sont les mots les plus proches de celui choisi comme point de départ[4]. Les nouvelles liaisons ainsi créées ont été baptisées « proxémies », et si cette proxémie peut être souvent confondue avec la synonymie (car elle relie deux mots généralement considérés comme substituables), elle aboutit également à des relations au sein d'un même champ sémantique, comme entre les verbes *déshabiller* et *éplucher*. L'intérêt du travail de Bruno Gaume, sur le lexique verbal par exemple, est de révéler par ces relations des substitutions qui apparaissent en production chez des sujets dont le lexique mental n'est pas encore fixé[5].

[4] A ce point de vue, on peut considérer son travail comme analogue à celui de Brodda & Karlgren.
[5] La raison de cet inachèvement peut être normale (chez des petits enfants) ou anormale ; c'est l'une des applications de ce travail, qu'il serait trop long de détailler ici.

3. Enrichissement exogène

Ce type d'enrichissement consiste à développer un programme d'extraction de synonymes applicable à divers corpus ; à vrai dire, la finalité de ce travail n'est pas forcément l'enrichissement d'un dictionnaire existant, mais plus souvent la production d'un dictionnaire ex nihilo. En outre, les principes mis en œuvre vont notablement différer selon qu'ils s'appliqueront à un corpus monolingue ou à un corpus multilingue.

Dans le cas du traitement d'un corpus monolingue, on commence généralement par analyser syntaxiquement le corpus en question, avant d'effectuer une analyse distributionnelle sur les unités repérées. L'idée qui sous-tend cette méthode est que les unités qui partagent des contextes distributionnels semblables sont sémantiquement proches. Ainsi, la mesure des similarités distributionnelles permet par la suite de faire apparaître les unités liées sémantiquement, comme l'a fait Didier Bourigault [Bourigault, 2002] avec un corpus de 10 années du journal « Le Monde » ou bien avec les textes des romans provenant de la base Frantext. Cependant, comme il l'a lui même montré [Bourigault et Galy, 2005], cette méthode rapproche assez peu d'unités synonymes, et quand bien même on appliquerait un filtrage catégoriel sur les résultats, les « voisins » obtenus compteraient parmi eux de nombreux antonymes, hyponymes ou hypernonymes.

La méthode que nous détaillerons ici améliore grandement la précision des résultats, grâce au choix d'un corpus multilingue aligné. Cette fois, l'idée sous-jacente est que si deux mots sont souvent traduits de la même manière dans de nombreuses langues, il y a une forte probabilité pour qu'ils soient synonymes. Ainsi, en utilisant les traductions, on trouve moins de mots apparentés parmi les unités similaires, car typiquement la traduction ne s'étend pas aux hyperonymes, (co)hyponymes ou antonymes. Par exemple, les mots « vin », « boisson » et « bière » ne se traduisent pas avec le même mot dans une autre langue. En outre, comme nous l'avons montré dans une étude précédente qui concernait le néerlandais, l'emploi de plusieurs langues donne de meilleurs résultats qu'un corpus bilingue, même dans le cas des langues les mieux apparentées [Van der Plas & Tiedemann, 2006].

Dans notre démarche, nous utilisons le corpus Europarl dont les textes proviennent des actes du Parlement Européen en 11 langues différentes[6], et dont nous avons tiré les traductions par des techniques dérivées de la traduction statistique automatique (outil open-source GIZA++). Plus précisément, les textes sont tout d'abord alignés phrase à phrase selon les techniques développées par Gale et Church [Gale & Church, 1993], puis mot à mot par GIZA++. Chaque mot se trouve ainsi pourvu de ses traductions dans les 10 autres langues avec leurs fréquences, ce qui constitue son vecteur caractéristique. Les similarités individuelles entre les mots sont ensuite calculées par comparaison de ces vecteurs, en tenant compte des fréquences des traductions, puisque l'indice prend en compte l'Information Mutuelle[7]. Finalement nous pouvons, pour les mots de la langue cible (ici le français), obtenir une liste de mots sémantiquement proches dont chacun est pourvu d'une similarité comprise entre 0 et 1. Un filtrage catégoriel est appliqué aux listes de synonymes proposés avant d'évaluer les résultats, comme dans le cas du corpus monolingue[8].

Les résultats pour le néerlandais ayant été satisfaisants [Van der Plas & Tiedemann, op. cit.], nous avons renouvelé l'opération avec la langue française, en changeant la référence d'évaluation des synonymes proposés en raison de certaines difficultés rencontrées avec la version néerlandaise de EuroWordnet. Pour le français, nous avons donc choisi le DES du

[6] Précisons qu'il s'agit des débats qui ont lieu au Parlement Européen, et non des textes issus de la Commission Européenne dont le technolecte n'est pas suffisamment riche.

[7] Pour plus de détails, on pourra se reporter à [Van der Plas & Tiedemann, 2006].

[8] Le filtrage élimine comme dans l'autre méthode certains résultats erronés, mais dont l'origine réside dans les problèmes d'alignement.

CRISCO. L'évaluation des résultats montre que les synonymes proposés par le système peuvent atteindre une précision double et un rappel triple de ceux obtenus avec un corpus monolingue, si l'on considère le dictionnaire des synonymes comme référence. Mais l'autre intérêt de la méthode, sur lequel nous voulons insister ici, c'est que ce traitement automatique fait apparaître des paires de synonymes qui devraient en principe se trouver dans le dictionnaire. Ainsi, les paires « affection – pathologie » ou « documentaire – reportage » sont clairement mises en évidence alors qu'elles sont absentes du dictionnaire. Par exemple, pour des similarités entre termes supérieures à 0,3, environ les deux tiers des relations proposées sont déjà présentes dans le dictionnaire ; mais parmi les relations qui en sont absentes et que le processus de traitement propose, un tiers forment des paires synonymiques parfaitement acceptables. La précision observée passe ainsi de 67 % à 77 %, et le dictionnaire peut par cette voie s'enrichir de nouvelles relations.

En outre l'apport des résultats ne se limite pas là, puisque le système produit aussi des paires dont les membres ne figurent pas forcément tous les deux parmi les entrées du dictionnaire. Par exemple, la détection de couples comme « adaptation – *reformulation* », « affaiblissement – *fragilisation* » ou « diversité – *pluralisme* » introduisent dans le dictionnaire les mots placés ici en italique. Cette faculté de détection prouve que la lexicographie synonymique trouve là un intérêt non négligeable.

Cela dit, deux problèmes liés, l'un d'ordre technique, l'autre d'ordre linguistique demeurent en suspens. Tout d'abord, les méthodes d'alignement statistique ne sont pas toujours à même de réaliser des appariements obéissant à l'organisation des énoncés ; c'est la raison pour laquelle nous avons appliqué à nos résultats un filtrage catégoriel pour mettre à l'écart des paires comme « majorité – majoritairement », celle-ci résultant bien sûr d'un alignement incorrect de la locution « en majorité ». Nous avons commencé à travailler sur ce problème en dotant le système d'alignement d'un dictionnaire de locutions, ce qui améliore grandement les résultats. La présence de ces paires hétérogènes (au point de vue catégoriel) révèle aussi un second problème qui pourrait se définir d'une manière globale par « la question des paraphrases ». En effet, le premier synonyme proposé par exemple pour « ville » est l'adjectif « urbain », et provient des difficultés d'alignement d'un mot sur des paraphrases comme « milieu urbain », « zone urbaine » ou « domaine urbain » ; mais il est relativement difficile de répertorier les paraphrases qui forment un ensemble ouvert.

Enfin, le fait que le rappel n'atteigne pas 50 % est un peu décevant, mais s'explique d'une part par le fait qu'une substitution d'un mot par un de ses synonymes va parfois s'accompagner d'un changement de niveau de langue, et que ce changement va aussi se retrouver dans les traductions. Par exemple, « pompe » est synonyme familier de « chaussure » selon notre dictionnaire, mais dans notre corpus ce mot n'est jamais utilisé dans ce sens, mais dans celui de l'appareil de pompage. Cela nous amène à l'autre explication de la faiblesse du rappel : la richesse du corpus ; il est évident que les débats entre parlementaires européens se cantonnent dans un niveau de langue assez soutenu, malgré la diversité des sujets abordés et la variété des avis formulés, et que cette forme de discours n'atteint pas la richesse d'un corpus journalistique de 10 années. Il est donc nécessaire pour nous de poursuivre ce travail par une mise à l'épreuve avec un corpus plus vaste et plus varié ; cette expérience de variété a débuté avec les sous-titrage des films, mais nous ne disposons pas à l'heure actuelle des sous-titres en français.

Conclusion : vers un dictionnaire des substituts ?

Comme nous l'avons vu, les traitements automatiques, qu'ils soient endogènes ou exogènes, constituent un jeu d'instruments intéressants pour le contrôle et surtout l'enrichissement des dictionnaires de synonymes. En outre, les perspectives qui apparaissent lors du traitement des corpus multilingues alignés laissent entrevoir des possibilités d'évolution pour cette sorte de dictionnaire. En effet, le traitement des unités complexes, et l'évolution des méthodes multilingues vers un alignement « fonctionnel » et non plus mot à mot, permettraient d'accéder à la construction de dictionnaires des substituts (ou de paraphrases) qui seraient d'un grand intérêt pour les apprenants d'une langue étrangère.

Bibliographie

BOURIGAULT Didier (2002) : « UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus », *Actes TALN 2002*, Nancy.

BOURIGAULT Didier & GALY, Edith (2005) « Analyse distributionnelle de corpus de langue générale et synonymie », Lorient, *Actes JLC 2005*.

BRODDA Benny & KARLGREN Hans (1969) : « Synonyms and synonyms of synonyms », *SMIL*, 5, pp. 3-17. Stockholm.

GALE, W. & CHURCH, K. (1993) : « A program for aligning sentences in bilingual corpora » In *Computational Linguistics*, 19(1).

GAUME Bruno (2006) : « Cartographier la forme du sens dans les petits mondes Lexicaux », *Actes JADT 2006*, Besançon.

HABERT Benoît (2005), *Instruments et ressources électroniques pour le français*, Paris, Ophrys.

KAHLMANN André (1975), *Traitement automatique d'un dictionnaire de synonymes*, Stockholm, Université de Stockholm.

LE FUR Dominique et al. (2006), *Dictionnaire des combinaisons de mots*, Paris, Editions le Robert.

MANGUIN Jean-Luc (2004) : « Transitivité partielle de la synonymie : application aux dictionnaires de synonymes », *CORELA – Cognition, Représentation, Langage*, Vol 2, n° 2.

PLOUX Sabine (1997). « Modélisation et traitement informatique de la synonymie ». *Linguisticae Investigationes*, XXI (1), Amsterdam, John Benjamins.

PRUVOST Jean (2006), *Les dictionnaires français outils d'une langue et d'une culture*, Paris, Ophrys.

QUEMADA Bernard (1968), *Les Dictionnaires du français moderne (1539-1863). Étude sur leur histoire, leurs types et leurs méthodes*, Paris, Didier.

VAN DER PLAS Lonneke & TIEDEMANN Jörg (2006), « Finding synonyms using automatic word alignment and measures of distributional similarity » *Actes de ACL/Coling 2006*, Sydney.

VICTORRI Bernard & FUCHS Catherine (1996), *La polysémie : une construction dynamique du sens*, Paris, Hermès.