HAL
open science

Identification de termes flous et génériques dans la documentation technique : expérimentation avec l'analyse distributionnelle automatique

Émilie Merdy, Juyeon Kang, Ludovic Tanguy

▶ **To cite this version:**

Émilie Merdy, Juyeon Kang, Ludovic Tanguy. Identification de termes flous et génériques dans la documentation technique : expérimentation avec l'analyse distributionnelle automatique. Atelier "Risque et TAL" dans le cadre de la conférence TALN, Jul 2016, Paris, France. ⟨hal-01365926⟩

HAL Id: hal-01365926

https://hal.science/hal-01365926v1

Submitted on 13 Sep 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Autorisation HAL

HAL Authorization

Identification de termes flous et génériques dans la documentation technique : expérimentation avec l'analyse distributionnelle automatique

Émilie Merdy (1 et 2) Juyeon Kang (1) Ludovic Tanguy (2)
(1) Prometil, Toulouse
(2) CLLE-ERSS : CNRS & Université de Toulouse, France
e.merdy@prometil.com, j.kang@prometil.com, tanguy@univ-tlse2.fr

RÉSUMÉ

Cette étude se place dans le cadre du développement des ressources linguistiques utilisées par un système de vérification automatique de documentations techniques comme les spécifications. Notre objectif est d'étendre semi-automatiquement des classes de termes intrinsèquement flous ainsi que des termes génériques afin d'améliorer le système de détection de passages ambigus reconnus comme des facteurs de risque. Nous mesurons et comparons l'efficacité de méthodes d'analyse distributionnelle automatiques en comparant les résultats obtenus sur des corpus de taille et de degré de spécialisation variables pour une liste réduite de termes *amorces*. Nous montrons que si un corpus de taille trop réduite est inutilisable, son extension automatique par des documents similaires donne des résultats complémentaires à ceux que produit l'analyse distributionnelle sur de gros corpus génériques.

ABSTRACT

Identification of fuzzy and underspecified terms in technical documents : an experiment with distributional semantics

This study takes place in the framework of the development of linguistic resources used by an automatic verification system of technical documents like specifications. Our objective is to enlarge semi-automatically the classes of intrinsically fuzzy terms along with generic terms in order to improve the steps of identifying ambiguous elements of the system such as factors of risk. We measure and compare the efficiency of the methods of automatic distributional analysis by considering obtained results from corpora of different sizes and specialization degrees by priming from a reduced list of prime terms. We show that if a corpus of too limited size is not useable, its automatic extension by similar documents produces results that can be completed by those obtained from distributional analysis on large generic corpora.

MOTS-CLÉS : ambiguïté, documents techniques, risque industriel, Analyse Distributionnelle Automatique (ADA), ressources lexicales.

KEYWORDS: ambiguity, technical documents, industrial risk, Automatic Distributional Analysis (ADA), lexical resources.

1 Introduction

Dans un contexte industriel où toute conception complexe doit être formalisée par écrit et où ces traces écrites peuvent désigner une entité fautive en cas d'échec ou d'accident, les spécifications techniques

se multiplient. À l'échelle d'une organisation, le temps passé à rédiger et corriger ces documents représente un investissement très important et la qualité obtenue est très variable d'un projet à l'autre, notamment en fonction de différents facteurs relatifs aux rédacteurs (expertise, fatigue, ...). Différents standards cohabitent à l'intérieur d'un domaine et à l'intérieur d'une organisation pour harmoniser ces documents et certains outils permettent d'assister la relecture pour en assurer la qualité. Parmi ces outils, Semios (Kang & Merdy, 2015), branche industrialisée du projet *Lelie* présentée dans Kang & Saint-Dizier (2015), détecte l'ambiguïté dans les documents techniques tels que les exigences et les procédures. Cette détection repose sur un ensemble de ressources lexicales et de patrons morpho-syntaxiques créés manuellement à partir de connaissances du domaine (experts et documents relatifs au domaine) et d'observations des contextes d'apparition des termes contenus dans les ressources lexicales.

Contexte

Les documents techniques, notamment les spécifications (exigences), les procédures (manuels d'utilisateur, les cas de test, les consignes de sécurité, les *do-lists* utilisées en aéronautique, etc.) forment un genre spécifique de la rédaction. Un haut niveau de qualité est exigé pour que ces documents soient facilement compréhensibles par différents lecteurs et que ceux-ci ne soient pas confrontés à des cas où différentes interprétations sont possibles, autrement dit à des cas ambigus.
Prenons comme exemple les spécifications : il s'agit de documents dans lesquels s'expriment les besoins fonctionnels, i.e. ce qu'un futur système doit faire et ce que les utilisateurs finaux attendent. L'ambiguïté dans les spécifications peut conduire le concepteur à prendre une décision qui n'est pas cohérente avec les attentes que le rédacteur pensait avoir exprimées. La concurrence de plusieurs interprétations peut provoquer des risques non seulement techniques (ex. mauvais fonctionnements d'un système) et économiques (surcoût de développement) mais surtout humains, écologiques et sociétaux en cas d'accident, etc. Voir notamment le rapport *Chaos Manifesto* (The Standish Group, 2013) qui répertorie les causes d'échecs des projets industriels et place les exigences mal gérées parmi les principales.

L'ambiguïté peut être de plusieurs types et prendre plusieurs formes (Zhang, 1998), et est bien entendu la cible principale lorsqu'on cherche à améliorer la qualité des spécifications (Tjong, 2008). Des normes de rédaction et guides de bonne pratique (INCOSE, 2012; Hull *et al.*, 2011) répertorient différents ensembles d'expressions susceptibles d'accroître l'ambiguïté dans les documents techniques. Si le recouvrement entre les cas ambigus n'est pas identique d'un standard à l'autre, certains phénomènes sont systématiquement présentés, à l'image de l'exemple ci-après.

L'ambiguïté peut venir de termes intrinsèquement ambigus, tels que "*approximately*", "*nearly*" ou "*appropriate*" comme dans cet exemple tiré de notre corpus d'exigences (décrit dans la section suivante) : *Thermal insulation shall be* appropriately *installed to minimize retention of liquids.* Ces termes et leurs synonymes font partie d'une classe limitée, majoritairement composée d'adjectifs et d'adverbes, non liée à un domaine spécifique mais également équivalente dans la langue générale. La constitution d'un lexique de ces termes demande donc un investissement raisonnable puisque la généricité de la ressource la rend utilisable dans tous les traitements sémantiques.

Les autres cas classiques d'ambiguïté lexicale concernent les termes nominaux ou verbaux sous-spécifiés, tels que "*system*", "*component*", "*manage*", "*operate*", etc. comme dans cet autre exemple de notre corpus : *The* components *shall be designed to* operate *during the operating life [...].*

Une exigence devant être intelligible isolément, ce type de phénomène anaphorique est donc à proscrire. À ce stade, les termes sous-spécifiés à envisager sont très nombreux, et un lexique contenant tous les hyperonymes de très haut niveau (*system*, *element*, ...) d'un ensemble de documents nécessite

d'être construit semi-automatiquement. De plus, les hyponymes de ces termes sont susceptibles d'être eux aussi ambigus dans le cas où il s'agit de syntagmes nominaux complexes parfois incomplets :

"*The (*system | interface | XX interface*) shall display the last command.*"

- *system* -> Hyperonyme de très haut niveau - terme sous-spécifié
- *interface* -> Hyponyme générique - terme sous-spécifié
- *XX interface* -> Hyponyme spécifié

La situation se complique lorsque certains termes ne sont pertinents qu'au niveau du domaine d'activité correspondant aux spécifications analysées. Par exemple, dans cet extrait du corpus d'exigences décrit dans la section suivante :
The XX Controller shall request the APU *as the bleed air source by sending signal YY.*

Le terme *APU* (pour *Auxiliary power unit* ou groupe auxiliaire de puissance) est en fait un terme sous-spécifié pouvant faire référence à plusieurs composants, ce qui est une source d'ambiguïté du même ordre que l'exemple précédent. La différence vient de la spécialisation du terme *APU*, qui est utilisé comme terme générique dans la conception de véhicules. Une analyse manuelle ne peut couvrir l'ensemble des contextes possibles sans avoir les connaissances du domaine et impliquer un coût très important pour que les linguistes adaptent l'outil après s'être familiarisé avec des documents destinés à des experts d'un domaine précis.

Il existe des méthodes et des outils qui aident à détecter des cas d'ambiguïté dans les documents techniques et alertent les rédacteurs (Zowghi & Coulin, 2005; De Gea *et al.*, 2012). Identifier des éléments ambigus dans les spécifications et réduire leur nombre avant de passer aux étapes de test et production peut avoir un impact important sur l'issue d'un projet industriel en termes de coûts et de temps. Cependant, il s'avère que l'ambiguïté est difficilement définissable de façon absolue puisque le caractère ambigu d'un élément (lexical, syntaxique ou sémantique) est fortement influencé, voire déterminé, par son contexte.
a) *The Archiving units shall be able to archive the* ***maximum*** *amount of data.*
b) *The* ***maximum*** *pressure loads at the standard operating temperature shall be 6.*
Dans l'exemple a), l'adjectif ***maximum*** est ambigu dans le sens où la valeur de "***maximum*** *amount*" n'est pas quantifiable. Ceci introduit plusieurs interprétations possibles. Par contre, dans b), il n'est pas ambigu car il s'agit d'une définition de la valeur de "***maximum*** *pressure loads*".

Ainsi, la seule détection d'un terme potentiellement ambigu ne suffit pas, et les outils comme Semios font intervenir des techniques plus complexes, comme des patrons lexico-syntaxiques, des systèmes de pondération et de faisceaux d'indices (que nous ne détaillerons pas ici). Dans tous les cas cependant, ces techniques se basent sur des classes de termes (génériques mais aussi adaptées aux domaines d'application voire aux clients et/ou aux projets), si bien qu'un manque de couverture lexicale provoque systématiquement du silence.

Nous allons donc dans la suite de cet article présenter un dispositif expérimental visant à construire ces ressources lexicales, en utilisant des techniques d'analyse distributionnelle. Plus précisément, nous allons voir dans quelles limites il est possible de partir d'un ensemble réduit de termes amorces (marques de flou et termes génériques) pour élargir la couverture de la base de données lexicale. Comme on l'a vu, ces ressources relèvent de différents degrés de spécialisation, et nous envisagerons donc différents types de corpus sur lesquels effectuer ces acquisitions par des méthodes automatiques. L'objectif ici est multiple : déterminer la nature et le volume des corpus à partir desquels sont détectables des relations sémantiques pertinentes, étendre un lexique d'hyperonymes de très haut niveau (axe syntagmatique) et se servir ce lexique pour construire des lexiques spécifiques à partir

des relations paradigmatiques détectées entre les hyperonymes et leurs voisins distributionnels.

Nous commençons par décrire les données dans la section suivante : le corpus d'exigences qui servira de banc de test, ainsi qu'un sous-ensemble des ressources lexicales de Semios. Puis nous présentons le dispositif expérimental mettant en jeu des méthodes d'analyse distributionnelle automatique (ADA) et les résultats obtenus.

2 Description des données

Dans cette partie, nous décrivons les différents corpus et ressources utilisées dans notre étude. Notre corpus de départ se compose de 5 spécifications issues d'un même projet, fournies par un partenaire industriel et anonymisées pour des raisons de confidentialité. Une spécification est un document relativement autonome, contenant un ensemble d'exigences ainsi que d'autres informations textuelles comme une justification concernant une exigence, et des figures (graphiques, schémas, etc.). Dans cette étude, les éléments non textuels ne sont pas traités. Ces 5 spécifications contiennent 5 186 balises uniques d'exigences rédigées en anglais, ce qui représente plus de 200 000 tokens (4 378 types). Toutes ces spécifications traitent de la conception de moteurs mais ne sont pas au même stade de maturité de rédaction (une spécification validée et les quatre autres sont à un stade intermédiaire, i.e. elles nécessitent des corrections et certaines sections sont incomplètes). La taille réduite de ce corpus ainsi que sa pauvreté lexicale sont des caractéristiques intrinsèques des corpus techniques spécialisés tout autant qu'une limitation en ce qui concerne les fondements statistiques sur lesquels reposent des techniques comme l'analyse distributionnelle.

Pour augmenter la couverture lexicale et syntaxique de notre corpus de spécifications, nous avons donc construit un corpus grâce à BootCaT (Baroni & Bernardini, 2004). Cette chaîne de traitement construit semi-automatiquement un corpus à partir d'amorces lexicales complexes qui servent de requêtes à un moteur de recherche (*Bing*) pour aspirer des pages web. Différents paramètres permettent de limiter le bruit, comme l'exclusion de termes et d'URL, tandis que d'autres augmentent le nombre de pages ramenées jusqu'à un certain seuil (nombre de requêtes par amorce, combinaison des amorces pour créer des requêtes plus complexes, etc.). Pour orienter la spécialisation de son contenu, nous avons sélectionné ce corpus web en plusieurs étapes pour évaluer la pertinence des résultats en fonction des paramètres cités plus haut, ainsi qu'en fonction des fréquences relatives des termes en corpus avec le concordancier AntConc (Anthony, 2005), d'une analyse terminologique réalisée avec YaTeA (Hamon, 2012) et d'un ensemble de termes observés en corpus et jugés spécifiques. La liste d'amorces est constituée de 51 termes complexes, principalement nominaux (48) tels que "*cockpit interface associated command*" et "*torque motor control optimization*" mais également prépositionnels (3) comme "*per flight hour*". Notre volonté est que ce corpus soit suffisamment similaire au corpus d'exigences en termes de contenu lexical et de contenu spécialisé. Cette validation souffre cependant de critères stables et mesurables pour estimer, même approximativement, les frontières ou les chevauchements de domaines, et ce corpus sera amené à évoluer à mesure que des amorces plus pertinentes seront identifiées.

À des fins de comparaison, nous avons également utilisé un corpus générique de grande taille, à savoir un sous-ensemble de 200 millions de mots du corpus UKWaC (Ferraresi *et al.*, 2008) supposé être représentatif de la langue anglaise générale telle qu'elle est rencontrée sur internet. Le corpus entier compte 2 milliards de mots mais à l'échelle de nos corpus techniques (200 000 et 2 millions de mots) le rapport de taille semble suffisant, tout en étant plus facilement manipulable.

Ressources constituées manuellement

Ici, nous présentons les ressources lexicales composées des termes ambigus sur lesquelles repose

Semios. Nous utilisons ces lexiques comme sources partielles (combinées à des recommandations du standard IEEE/ISO 29148-2011) pour sélectionner des amorces lexicales connues pour leur potentiel ambigu dans les exigences. Dans un second temps, ces ressources servent également de base pour estimer la couverture et l'apport de la méthode distributionnelle appliquée dans cette étude telle qu'elle est décrite dans la section 3. Des lexiques de termes flous ainsi que des lexiques de termes génériques relatifs aux domaines aéronautique, automobile et naval ont été développés par des linguistes dans le cadre de Semios. La construction de ces ressources peut être décrite en trois étapes :

1) Analyse des standards IEEE (notamment, IEEE/ISO 29148-2011), des guidelines (INCOSE, 2012) et IREB (Pohl & Rupp., 2011), et des principes des langues contrôlées (ASD, 2013). Dans ces sources, nous observons les bonnes pratiques de rédaction des documents techniques et les principes définis couvrent la syntaxe, la sémantique, le lexique et le style que les rédacteurs doivent respecter. Nous y trouvons la liste de base des termes qui sont trop génériques pour que l'objet auquel ils font référence soit non-ambigu sans précision : "*the system*", *the software*", "*provide a field for*", "*operate at a power level*", "*almost always*", "*as applicable*" etc.

2) Expériences industrielles permettant d'identifier des termes qui amènent de l'ambiguïté dans les documents techniques. Ce travail se fait par l'analyse manuelle des spécifications et des guides de rédaction des industries. La fouille des spécifications est une source majeure pour enrichir le lexique de base. Là, les nouveaux termes génériques sont identifiés, comme "*malfunction*" ou "*undesirable effects*" qui, dans l'exemple suivant, ne précisent pas les états jugés problématiques :
"*X and Y systems/equipments shall operate without degradation of performance malfunction or undesirable effects during [...]*"

3) Étude spécifique au domaine pour la construction des termes métier. Pour un outil industriel, il est primordial de réaliser cette étude d'adaptation au domaine spécialisé et par la suite, de construire un lexique des termes métier. Ces termes métier font souvent partie des sources des faux positifs, par exemple, le terme "*normal mode*" qui est candidat au terme ambigu dans un contexte générique est un concept précis pour les experts du domaine dans un contexte aéronautique. Ce dernier point nécessite une observation fine du contexte et une prise en compte du domaine pour distinguer les cas ambigus des non-ambigus mais dans cette étude nous nous focalisons sur les termes potentiellement problématiques par nature et ne développons pas la prise en compte du contexte.

3 Une application de l'analyse distributionnelle automatique

Cette section présente le dispositif expérimental déployé pour évaluer l'apport des méthodes d'analyse distributionnelle.

Principes
Les méthodes de l'ADA font désormais partie des outils classiques utilisés dans le TALN pour un ensemble d'approches et d'applications visant ou impliquant la sémantique lexicale. Quelle que soit la technologie utilisée, ces techniques reposent toutes sur l'hypothèse distributionnelle d'Harris (Harris, 1954) qui exprime que deux mots qui ont un comportement similaire en corpus (partagent un ensemble de contextes) entretiennent une forme de similarité sémantique.

Si à l'origine ces méthodes visaient l'exploration de corpus de langue de spécialité, elles se sont depuis étendues à de grands corpus génériques supposés représentatifs du fonctionnement général du lexique d'une langue. L'application la plus directe dans les deux cas est l'identification non supervisée de relations entre éléments du lexique, aboutissant suivant les cas à des réseaux de similarités, des thésaurus ou des classes lexicales (Fabre & Lenci, 2015).

À ce stade de développement de ces techniques, on trouve aisément des ressources génériques comme les mémoires distributionnelles construites sur de gros corpus (Baroni & Lenci, 2010) mais aussi des boîtes à outils prêtes à l'emploi, la plus répandue actuellement étant Word2vec (Mikolov *et al.*, 2013). La forme la plus simple de ces ressources et des sorties de ces outils est un ensemble de relations valuées qui traduisent la proximité distributionnelle entre deux unités lexicales, calculées sur la base d'un corpus. La nature des relations lexicales mises au jour par ces méthodes reste sous-spécifiée, mais on sait qu'elles recouvrent l'ensemble des relations classiques : synonymie, hyponymie, co-hyponymie et antonymie notamment.

Application à nos corpus

Nous avons utilisé la boîte à outils Word2Vec (Mikolov *et al.*, 2013) pour construire des modèles distributionnels à partir des trois corpus décrits précédemment. Plus précisément, chaque corpus a été segmenté, étiqueté et lemmatisé par TreeTagger pour isoler et normaliser les mots en couples lemme_catégorie (*works* et *worked* deviennent tous les deux *work_V*). Les paramètres suivants ont été utilisés pour construire le modèle distributionnel : modèle skip-gram pour représenter les contextes avec une fenêtre de 6 mots, réduction à 200 dimensions, méthode *hierarchical softmax*. À l'issue de ce calcul, nous disposons donc d'une matrice qui associe à chaque mot du corpus un vecteur (de longueur 200) indiquant ses coordonnées dans l'espace vectoriel.

Pour interroger ces bases, nous calculons simplement pour un mot donné la liste des 60 mots du corpus qui lui sont le plus proches (également appelés voisins distributionnels) suivant le modèle distributionnel en se basant sur une similarité cosinus entre les vecteurs correspondants aux mots. Puisque notre corpus est étiqueté, nous ajoutons un filtrage supplémentaire en ne retenant que les voisins qui ont la même catégorie grammaticale que le mot de départ.

À titre d'exemple, en utilisant le modèle distributionnel construit sur UKWaC, les premiers voisins adjectivaux de "*approximate"* sont : "*exact*", "*estimated*", "*actual*", "*recommended*" et "*related*". On peut y voir pêle-mêle des synonymes (*estimated*) et des termes liés comme *recommended* dont le statut pour notre problématique est similaire à celui du pivot *approximate*, mais aussi des antonymes comme *exact* qui eux ont un statut différent.

Nous avons sélectionné 16 termes désignés comme étant potentiellement flous soit par IEEE/ISO 29148-2011 soit par nos observations de leur comportement dans le corpus d'exigences. Il se répartissent en quatre catégories grammaticales : 8 adjectifs ("*easy*", "*appropriate*", "*best*", "*large*", "*most*", "*normal*", "*effective*" et "*significant*"), 3 adverbes ("*about*", "*regularly*" et "*almost*") et 5 noms ("*system*", "*malfunction*", "*component*", "*element*" et "*software*"). Parmi les 16 termes sélectionnés, les 3 adverbes et 8 adjectifs des termes porteurs d'ambiguïté quel que soit le domaine dans lequel ils apparaissent. Les 5 noms posent quant à eux des problèmes d'interprétation uniquement s'ils sont sous-spécifiés, c'est-à-dire qu'il s'agit d'hypéronymes de très haut niveau qui ont besoin d'être qualifiés par des modifieurs pour être distingués.

Les résultats sont quantitativement très inégaux. En effet, après filtrage, "*about*" ne ramène aucun voisin issu du corpus d'exigences - à cause d'une fréquence trop basse - alors que 5 voisins sont identifiés dans le corpus technique de pages web, et 21 sont issus de UKWaC. Dans quelques cas, les voisins distributionnels du corpus d'exigences sont plus nombreux que ceux du corpus web spécialisé, ce qui peut sembler surprenant, mais la précision de l'étiquetage syntaxique et le filtrage sur les catégories explique partiellement cette répartition finale. UKWaC est systématiquement plus prolifique en termes de voisins conservés après filtrage, ce qui s'explique logiquement par sa taille.

Sur le plan qualitatif, le tableau 1 présente la liste des premiers voisins ramenés pour le mot *malfunc-*

tion sur chacun des 3 corpus.

	Corpus d'exigences	Corpus web spécialisé	Corpus UKWaC
Rang 1	**degradation**	indoor	**harm**
Rang 2	fluid	**abnormality**	**interruption**
Rang 3	do-160g	thermistor	**delay**
Rang 4	**damage**	outdoor	**trouble**
Rang 5	service	**failure**	mce
Total au rang 10	2/10	5/10	5/10
Total global	4/31	9/31	11/49

TABLE 1 – Répartition des voisins distributionnels de "*malfunction*" dans les 3 corpus

Comme on peut l'observer dans le tableau 1, les relations sémantiques entre "*malfunction*" et ses voisins distributionnels sont complémentaires d'un corpus à l'autre. Une fois les erreurs d'étiquetage écartées, la couverture lexicale se répartit sur les trois corpus sans présenter de redondance. Sans surprise, le corpus web spécialisé fournit plus de réponses pertinentes que le corpus d'exigences, malgré un bruit imputable à l'étiquetage. Si le nombre de réponses au rang 5 font de UKWaC un candidat idéal, les voisins pertinent n'augmentent que difficilement sur la fin de la liste.

Évaluation et discussion

L'ensemble des voisins rapportés par cette méthode a été évalué par 3 juges présentant un niveau d'expertise moyen, avancé et très avancé vis-à-vis des exigences du domaine observé, et ayant tous les trois une très bonne connaissance du système Semios. La consigne pour l'évaluation était formulée comme suit : est-ce que ce voisin présente une ambiguïté dans le champ sémantique du pivot et devrait figurer dans la même ressource lexicale ? Le taux d'accord observé deux à deux dépassant les 80% pour cette tâche, la réponse majoritaire a été sélectionnée. Le tableau 2 présente le résultat global sous la forme "voisins jugés pertinents/voisins ramenés".

Par exemple, parmi les voisins de *malfunction* visibles dans la table 1, les termes en gras sont ceux qui ont été considérés comme valides (*degradation*, *damage*, etc.). Les termes rejetés sont soit totalement inappropriés (mauvais étiquetage comme indoor ou mauvaise segmentation) soit sans lien sémantique probant (service). Si le corpus d'exigences fait émerger uniquement deux co-hyponymes du terme pivot, le corpus web spécialisé contient lui aussi deux co-hyponymes parmi les 5 premiers termes, différents toutefois de ceux issus du corpus d'exigences. Malgré une densité plus forte des premiers résultats (3 hyponymes "*interruption*", "*delay*" et "*trouble*" et 1 terme lié "*harm*" sur les 5 premiers voisins) et une productivité plus importante, au rang 10, le corpus UKWaC ne dépasse pas la performance quantitative du corpus web spécialisé et sur son total de 49 voisins potentiels, seulement 11 sont sémantiquement liés. Cependant, à part "*damage*" et "*failure*", il n'y a pas de recouvrement entre les différents corpus et les types de relations ne semblent pas être réparties de façon homogène, le corpus UKWaC présentant plus de liens d'hyperonymie ("*fault*", "*problem*", ...) tandis que le corpus web spécialisé contient plutôt des relations de même niveau du point de vue de la généricité ("*overheat*", "*disconnection*", "*rupture*", ...).

On observe globalement que le nombre de voisins pertinents, mais aussi le taux de pertinence lui-même croit rapidement avec la taille des corpus (rappelons qu'il y a un rapport de 1 à 10 et 1000 entre les trois). Le faible retour observé pour le corpus initial nous pousse à le considérer impropre à un usage par l'analyse distributionnelle automatique, du moins telle que nous l'avons pratiquée et sur les pivots étudiés. La distinction entre les deux autres corpus étudiés est par contre un peu plus complexe.

	Corpus d'exigences	Corpus web spécialisé	Corpus UKWaC
Adjectifs	6/33	39/126	151/260
Adverbes	0	8/21	27/71
Noms	15/192	24/175	50/256
Total	21/325	71/322	228/587

TABLE 2 – Répartition des voisins jugés pertinents

La productivité des adjectifs et adverbes flous est nettement plus importante dans un corpus de très grande taille sans restriction sur un domaine de spécialité (UKWaC). Cette conclusion est logique puisque ces termes n'ont pas de contrainte particulière liée au domaine, et bénéficie de la grande taille du corpus qui garantit à la fois une meilleure couverture lexicale et une plus grande efficacité des méthodes distributionnelles. De même, pratiquement aucun des voisins valides repérés pour ces pivots dans les corpus spécialisés n'est pas également ramené par UKWaC, ce qui nous conduit à considérer qu'un gros corpus générique est préférable pour étendre les classes lexicales autour de ces termes. La tendance en volume est la même pour les noms génériques, et là aussi la taille du corpus semble la principale explication pour la masse de noms voisins valides ramenés. Par contre un ensemble de différences qualitatives sont à considérer lorsque l'on compare les trois corpus. Ainsi, les voisins pertinents issus de UKWaC présentent majoritairement le même degré de généricité, et un terme a toujours plus de voisins génériques issus de UKWaC que du corpus spécialisé étendu. C'est le cas notamment de "*mechanism*", "*device*" ou "*tool*" pour le pivot "*system*". Le corpus web spécialisé fait par contre émerger des voisins hyponymes des pivots, et ces relations d'hyponymie couvrent différents degrés de spécificité : "*subsystem*" et "*unit*" sont hyponymes de "*system*". Le recouvrement des voisinages entre les différents corpus est faible, puisque seulement 9 termes se retrouvent dans deux des ressources, et aucun terme n'est commun aux trois, ce qui prouve que les différents corpus sont complémentaires en fonction du type de relation recherchée.

Si l'on se concentre sur les voisins plus spécifiques (hyponymes), par exemple pour le pivot "*element*", on trouve des différences notables entre les corpus. Les quatre termes trouvés dans le corpus web spécialisé sont des hyponymes de très bas niveau tels que "*sensor*", "*inhalator*", "*aftercoolers*" et"*pipe*". Ces termes renvoient à des pièces composant ou interagissant directement avec des moteurs, nous sommes donc précisément dans le domaine de nos exigences, i.e. la conception de moteurs. À l'inverse, les hyponymes identifiés dans le corpus UKWaC ("*function*", "*parameter*" et"*ingredient*") renvoient d'une part à un domaine connexe (informatique) ou à des notions extérieures au domaine visé, il est donc clair qu'ils sont moins spécifiques et a fortiori moins pertinents pour construire une ressource lexicale adaptée. Ces termes plus éloignés éloignent de fait les voisins plus spécifiques en empêchant la mise au jour de termes pertinents.

4 Conclusion

Nous avons mené une première expérience pour évaluer la pertinence des méthodes d'analyse distributionnelle automatique afin de construire et d'étendre des ressources lexicales pour détecter l'ambiguïté dans des documents techniques comme les exigences. En partant d'un corpus de très petite taille (200 000 mots), comme le sont généralement les corpus techniques spécialisés, nous avons construit par la méthode BootCat un corpus de taille intermédiaire (2 millions) que nous avons estimé circonscrit au même domaine que le premier. Nous avons enfin utilisé un corpus Web générique (200 millions de mots).

Sur la base de 16 mots-amorces (adjectifs et adverbes flous et noms sous-spécifiés) nous avons évalué les voisins distributionnels renvoyés par la même méthode (Word2vec) sur ces trois corpus. Il apparaît que le corpus initial est de trop petite taille pour fournir des résultats exploitables, quel que soit le terme de départ.

Pour ce qui concerne l'identification de termes flous génériques (adjectifs et adverbes), qui ne sont pas liés à un domaine précis, il apparaît clairement que le corpus UKWaC est le plus adapté, de par sa taille, puisque le phénomène visé est largement répandu. L'observation du voisinage sémantique des adverbes et adjectifs flous ("*around*", "*regularly*", ...) confirme que l'ADA, et plus précisément que la boîte à outils Word2Vec telle qu'elle est distribuée, est une technique idéale pour constituer semi-automatiquement des lexiques de termes issus de classes restreintes à partir de corpus de très grande taille, sans se limiter à un domaine.

Pour les termes sous-spécifiés, par contre, l'utilisation d'un corpus générique et du corpus intermédiaire semblent complémentaires. Si pour les termes de grande généricité le corpus UKWaC est là encore plus productif, dès que ce sont des hyponymes qui sont visés le corpus spécialisé apporte une plus-value évidente en garantissant des termes plus pertinents.

Comme sur les corpus de très grande taille et non-spécialisés, l'ADA permet de détecter des classes sémantiques (synonymes, antonymes, hyperonymes et co-hyponymes) à partir de documents techniques, à condition qu'ils dépassent un seuil critique en termes de richesses lexicales et syntaxiques. Si la taille minimale d'un corpus pour faire apparaître les voisins distributionnels de termes spécifiques n'est pas définissable dans l'absolu, l'expansion d'un corpus par des pages web obtenues à partir de requêtes composées de termes spécifiques complexes permet d'automatiser une partie de la tâche de constitution et complétion de ressources lexicales adaptées.

L'étude des noms génériques ("*system*", "*component*", ...) appuie elle aussi la pertinence de cette méthode pour identifier des termes spécifiques au domaine entretenant des relations d'hyponymie (et d'hyperonymie dans une moindre mesure), cependant dans ce cas il est nécessaire d'exploiter des corpus spécialisés. Pour aller plus loin, la mesure de certains critères tels qu'ils sont exposés dans (Condamines & Warnier, 2014) peut aider à qualifier les données exploitées pour évaluer la similarité de registre. Cependant, plusieurs domaines peuvent se croiser dans les documents spécialisés, comme c'est le cas dans notre corpus d'exigences qui relève principalement de la conception de moteurs mais également développement informatique. La difficulté à délimiter un domaine est notamment visible à travers les voisins distributionnels de "*system*" qui renvoient plutôt à un système mécanique dans les deux corpus spécialisés et à un système informatique dans le corpus UKWaC.

Perspectives

Cette étude préliminaire s'est donc limitée à l'observation de termes simples, mais l'adaptation de la méthode distributionnelle à l'observation du voisinage sémantique de termes complexes permettra de mener des analyses plus poussées. En effet, imaginons que dans le domaine informatique l'adjectif "*normal*" doive systématiquement être signalé comme un terme ambigu. Lors de l'analyse de documents issus du domaine aéronautique ou automobile, "*normal mode*" ou "*normal conditions*" ne doivent pas déclencher d'alertes puisqu'il s'agit, dans ce cadre spécifique, de termes faisant référence à des concepts précis et communs à tous les opérateurs qui les emploient. Sans l'intervention d'experts de ces domaines, seule une identification automatique de termes complexes spécifiques peut distinguer les emplois ambigus des emplois spécialisés au domaine de certains termes sémantiquement ambivalents.

Pour aller plus loin, nous mettons en place une méthode distributionnelle adaptée à la détection

d'associations syntagmatiques à travers une approche de l'ADA qui s'appuie sur les dépendances syntaxiques. De cette manière nous espérons détecter les syntagmes nominaux complexes dans le but d'assister la tâche d'identification des noms qui ont besoin de modifieurs distinctifs ("*system*", "*component*", ...). L'identification des compléments nominaux et prépositionnels de certains verbes fera également l'objet d'une étude ultérieure pour détecter une autre manifestation de l'ambiguïté liée à la sous-spécification. Ce phénomène se présente quand des verbes tels que "*to consider*", "*to enable*", "*to operate*", et "*to provide*" ne sont pas accompagnés de modifieurs qui les spécifient. Le seuil de la fenêtre graphique a été fixé à 6 mots pour construire le modèle distributionnel de ce corpus. Ce choix repose sur la volonté de détecter des associations privilégiées entre un verbe et son ou ses modifieur-s, malgré des distances variables entre les deux (conjonctions, énumérations, etc.). Bien que le modèle ait été construit selon l'approche *skipgram* et non *bag of words* de Word2Vec, les résultats de notre analyse n'ont pas été concluants : peu de rappel et bruit conséquent. Ce constat appuie les conclusions d'une expérimentation portant sur la comparaison de modèles distributionnels à partir d'un corpus composé de 22 000 lemmes (cinq fois plus que dans le nôtre) : la prise en compte des relations syntaxiques pour obtenir des voisinages distributionnels syntagmatiques est pertinente à condition de disposer d'informations suffisamment fines (Tanguy *et al.*, 2015). Les premières observations des voisinages considérés selon cette approche sont prometteuses pour l'identification de modifieurs de termes nominaux comme l'illustre la liste ordonnée des 10 premiers voisins de "*malfunction* : *route*, *benefit*, *entry*, *sign*, *country*, *hole*, *unit*, *facility*, *designer*, *group* et *gen*". Ce constat encourage à examiner les voisins distributionnels de verbes porteurs d'une ambiguïté similaire à celle des noms sous-spécifiques.

Références

ANTHONY L. (2005). Antconc : design and development of a freeware corpus analysis toolkit for the technical writing classroom. In *Professional Communication Conference, 2005. IPCC 2005. Proceedings. International*, p. 729–737 : IEEE.

ASD (2013). *AeroSpace and Defence Industries Association of Europe - Specification ASD-STE 100*. Rapport interne, Issue 6.

BARONI M. & BERNARDINI S. (2004). Bootcat : Bootstrapping corpora and terms from the web. In *LREC*.

BARONI M. & LENCI A. (2010). Distributional memory : A general framework for corpus-based semantics. *Computational Linguistics*, **36**(4), 673–721.

CONDAMINES A. & WARNIER M. (2014). Linguistic analysis of requirements of a space project and their conformity with the recommendations proposed by a controlled natural language. In *Controlled Natural Language*, p. 33–43. Springer.

DE GEA J. M. C., NICOLÁS J., ALEMÁN J. L. F., TOVAL A., EBERT C. & VIZCAÍNO A. (2012). Requirements engineering tools : Capabilities, survey and assessment. *Information and Software Technology*, **54**(10), 1142–1157.

FABRE C. & LENCI A. (2015). Distributional semantics today : Introduction to the special issue. *TAL*, **56**(2), 7–20.

FERRARESI A., ZANCHETTA E., BARONI M. & BERNARDINI S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, p. 47–54.

HAMON T. (2012). Acquisition terminologique pour identifier les mots clés d'articles scientifiques. *Actes du huitième DÉfi Fouille de Textes*, p. 25–32.

HARRIS Z. S. (1954). Distributional structure. *Word*, **10**(2–3), 146–162.

HULL E., JACKSON K. & DICK J. (2011). *Requirement engineering*. Springer.

INCOSE (2012). *Guide for writing requirements*. Rapport interne, International Council on Systems Engineering, requirements working group.

KANG J. & MERDY E. (2015). Semios : Relecteur automatique d'exigences pour une aide à la rédaction de spécifications. *Génie Logiciel*, **115**, 10–16.

KANG J. & SAINT-DIZIER P. (2015). Une expérience d'un déploiement industriel de Lelie : une relecture intelligente des exigences. In *Actes de INFORSID*.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR 2013*, p. 1–12.

POHL K. & RUPP. C. (2011). *Requirements Engineering Fundamentals*. O'Reilly.

TANGUY L., SAJOUS F. & HATHOUT N. (2015). Évaluation sur mesure de modèles distributionnels sur un corpus spécialisé : comparaison des approches par contextes syntaxiques et par fenêtres graphiques. *TAL*, **56**(2), 105–129.

THE STANDISH GROUP (2013). Chaos manifesto.

TJONG S. F. (2008). *Avoiding ambiguity in requirements specifications*. PhD thesis, University of Nottingham.

ZHANG Q. (1998). Fuzziness-vagueness-generality-ambiguity. *Journal of pragmatics*, **29**(1), 13–31.

ZOWGHI D. & COULIN C. (2005). Requirements elicitation : A survey of techniques, approaches, and tools. In A. AURUM & C. WOHLIN, Eds., *Engineering and Managing Software Requirements*. Springer.