# Identifier et catégoriser l'ambiguïté dans les spécifications techniques de conceptions de systèmes

Émilie Merdy
CLLE-ERSS : CNRS & Université de Toulouse, France
Prometil, Toulouse
e.merdy@prometil.com

RÉSUMÉ

Cette étude s'inscrit dans le cadre d'une thèse Cifre avec Prometil [1], une société qui commercialise un outil de détection automatique des erreurs dans les exigences, i.e. le contenu textuel des spécifications techniques. Il s'agit d'un travail de recherche dans la lignée des travaux en analyse de corpus menés par le laboratoire CLLE-ERSS qui s'intéresse aux corpus spécialisés. Dans le cadre de l'adaptation automatique des analyses sémantiques à de nouveaux domaines, nous étudions la détection automatique de l'ambiguïté - qu'elle soit syntaxique, sémantique ou lexicale - dans les exigences à partir de ressources lexicales spécifiques mais incomplètes. En parallèle, l'exploration des exigences, qui sont des données non-massives et porteuses de peu de variétés lexicale et syntaxique, doit permettre de mieux appréhender la spécificité linguistique de corpus techniques spécialisés pour enrichir semi-automatiquement des ressources lexicales adaptées.

ABSTRACT

**Identifying and classifying ambiguity in requirements**

This research falls within an industrial PhD for Prometil, a company that commercializes an automatic proofreader dedicated to requirements, i.e. textual content of specifications. It is a scientific research in collaboration with CLLE-ERSS, a laboratory specialized in corpus analyses. Within the context of the adaptation of semantic analysis, we study automatic ambiguity detection in requirements from specific but incomplete lexical resources. At the same time, the examination of requirements, documents that result in a low amount of data and poor lexical and syntactic variety, should offer a better understanding of ambiguity in specialized technical corpora to enrich specific lexical resources semi-automatically.

MOTS-CLÉS : Sémantique lexicale et distributionnelle, Ambiguïté, Adaptation de domaine, Ressources lexicales, Spécifications et exigences.

KEYWORDS: Lexical and distributional semantics, Ambiguity, Domain adaptation, Lexical resources, Specifications and requirements.

## 1 Introduction

Cette étude s'inscrit dans le cadre d'une thèse Cifre avec une société qui développe et commercialise un outil d'analyse automatique de la qualité des spécifications, ou plus précisément des exigences,

1. http://prometil.com/

i.e. le contenu textuel des spécifications. Rédigées et utilisées dans des secteurs industriels variés (conception mécanique, de sites internet, de logiciels, etc.), les spécifications sont des documents à la fois techniques et contractuels autour desquels s'organisent la conception et le développement de systèmes ou la mise en œuvre de projets complexes. En plus des informations textuelles, une spécification peut contenir des éléments non-textuels tels que des schémas, graphiques, captures d'écran, etc. mais nous n'en tenons pas compte dans cette étude en raison de l'incapacité de les traiter automatiquement.

Selon les guides de bonnes pratiques, une exigence est référencée par un identifiant unique et doit être une description courte, simple et atomique de ce qu'un système doit faire. Les exigences sont majoritairement rédigées en langage naturel (parfois contraint, à l'image des langages contrôlés comme évoqué dans la section 2) et peuvent être accompagnées d'un commentaire justifiant l'attente exprimée. Ces commentaires ne sont pas régis par des recommandations de rédaction aussi contraignantes que les exigences mais, en raison d'un contenu lexical similaire, nous traitons tout le contenu textuel apparaissant après un identifiant d'exigence jusqu'à la fin de la justification si celle-ci est présente. L'extrait suivant [2] est une exigence fonctionnelle :

"*Opposite X position shall be computed with the same X position logics based on opposite inputs Rationale : Opposite information logic is exactly the same with differents inputs only*"

Les spécifications assurent deux fonctions aussi importantes l'une que l'autre. D'une part il s'agit de documents rédigés et utilisés comme supports techniques de travail par les professionnels qui conçoivent et développent des systèmes (informatiques, aéronautiques, etc.). D'autre part, ces documents ont valeur d'accord officiel entre les parties prenantes en cas de problème (échec du projet ou accident). Pour ces deux raisons, la non-ambiguïté est fondamentale pour éviter les situations anormales et pour déterminer à quelle partie attribuer la responsabilité en cas d'échec ou d'accident. Pour détecter l'ambiguïté, *Semios*, parmi les quelques outils d'analyse automatique des spécifications sur le marché, repose sur des règles symboliques et des lexiques créés et complétés manuellement. Actuellement, l'adaptation de ces ressources pour un nouveau domaine demande un investissement extrêmement important pour tenir compte des spécificités lexicales et syntaxiques des documents à traiter afin d'optimiser la pertinence des analyses.

Le but du travail de thèse dans lequel s'inscrit cette étude est donc de définir une méthodologie robuste pour constituer semi-automatiquement des ressources lexicales adaptées à tout domaine technique spécialisé à partir d'exigences. Pour y parvenir, nous commençons par examiner l'ambiguïté telle qu'elle est appréhendée par les guides de rédactions dans la section 2. Ensuite nous décrivons dans la section 3 le corpus d'exigences issues des spécifications d'une société industrielle spécialisée dans la conception de moteurs d'aéronefs, puis dans la section 4 nous passons en revue les différentes formes de l'ambiguïté telles qu'elles peuvent être rencontrées en corpus et appréhendées automatiquement. Enfin, la section 5 expose les actions entreprises pour poursuivre cette étude.

# 2 Ambiguïté dans les exigences et outils et techniques pour la contrôler

## 2.1 L'ambiguïté selon les guides de rédaction de spécifications

Cette section donne un aperçu des cas d'ambiguïtés décrits par les guides de rédaction tels qu'ils sont rencontrés dans notre corpus pour en proposer une catégorisation linguistique.

2. Sauf indication contraire, les exemples présentés sont tous issus du corpus d'exigences, décrit dans la section 3

L'extrait suivant présente une exigence qui doit être réécrite parce qu'elle contient plusieurs problèmes, mis en évidence en caractères gras (seulement les plus évidents sont notés mais d'autres phénomènes mériteraient également d'être discutés).

"*A Travel Limiter,* ***also referred to as*** *Pneumatic Finger, shall* ***be integrated*** *as part of the XX body.* ***This device*** *shall* ***be*** *pneumatically* ***driven****, according to values of differential pressure between the pressurized area and ambient.* ***It*** *provides a mean of* ***limiting*** *cabin depressurization in the event of* ***failures*** *leading to XX opening and* ***failures*** *of the XX function.*"

Le fait qu'elle exprime plus d'une attente peut poser problème dans la phase de test. En effet, plutôt que de cocher "oui" ou "non" pour signifier qu'un test du système est cohérent avec les attentes de l'exigence, le test est ralenti pour expliciter les cas de figures qui contiennent plus d'une attente, voire annulé si le protocole n'admet aucun écart à la règle. L'usage de "*also referred to as*" est une tournure qui allonge la phrase pour ajouter une information qui pourrait être donnée en dehors de l'exigence. L'emploi d'une forme passive qui n'explicite pas l'agent ("*be driven*) est également jugée comme une source de risque, tout comme le recours au pronom "It", notamment en début de phrase, puisque plusieurs interprétations sont en concurrence. L'emploi du terme "failures" pose lui aussi pose problème car il est trop générique, il renvoie potentiellement à différentes formes d'échec qui n'auraient pas toutes les mêmes conséquences ou implications.

Plusieurs standards cohabitent pour harmoniser les spécifications, tels qu'INCOSE [3] (International Council on Systems Engineering, ingénierie des exigences de haut niveau) ou IEEE (*Institute of Electrical and Electronics Engineers*) pour les documents relatifs à la conception de systèmes dans les domaines de la technologie de l'information et du nucléaire. À l'échelle de chaque organisation, des guides de rédaction peuvent se substituer, partiellement ou complètement, aux standards du domaine.

Parmi les définitions qui établissent ce qu'est une exigence, la suivante est attribuée à (IEEE, 1998) :

"*Une déclaration qui identifie* ***une caractéristique ou une contrainte*** *opérationnelle, fonctionnelle ou de conception* ***relative à un produit ou à une procédure****, qui est non-ambiguë, testable ou mesurable, et nécessaire à l'acceptabilité du produit ou de la procédure"* [4] "

L'extrait ci-dessous est une version simplifiée du standard IEEE/ISO/IEC (2011) :

- *L'exigence doit être* ***non-ambiguë*** *: une seule interprétation est possible. L'exigence doit être exprimée simplement et facile à comprendre.*
- *L'exigence doit être* ***complète*** *et ne pas dépendre d'une autre exigence, ni d'un schéma (qui est toléré s'il permet d'expliciter un propos), parce qu'elle est mesurable.*
- *L'exigence doit être* ***atomique*** *et* ***unique****, c'est-à-dire qu'elle doit se situer à un niveau de description suffisamment fin pour ne pas définir plus d'une exigence et ne pas être redondante avec une autre exigence.*

Ces règles de rédaction reposent sur des considérations de différentes natures, parfois même de différents niveaux (statut vis-à-vis du système, du contenu sémantique de l'exigence, de la relation entre différentes exigences).

Parmi les caractéristiques partagées pour définir une exigence, les notions de non-ambiguïté, de possibilité de tester ainsi que de mesurer apparaissent. Ces critères sont également centraux dans les

3. Présentation générale d'INCOSE disponible à cette adresse : http://www.incose.org/

4. "A statement that identifies a product or process operational, functional, or design characteristic or constraint which is unambiguous, testable or measurable, and necessary for product or process acceptability

recommandations d'autres standards de rédaction diffusés en interne à l'échelle des sociétés.

Les exigences peuvent être fonctionnelles (i.e. définir ce que le système doit faire) ou non-fonctionnelles (i.e. définir les qualités attendues d'un système). Cette étude se concentre sur un ensemble de cinq spécifications contenant des exigences fonctionnelles destinées à la conception d'un système de motorisation pour des aéronefs.

Si tous les standards (à l'échelle d'un domaine ou d'une société) ne diffusent pas exactement les mêmes recommandations, une exigence jugée *bien-formée* pour l'un nécessitera peu de modifications pour être acceptée par un autre. Certains guides conseillent de ne rédiger que des "phrases simples", de ne "pas employer de conjonctions" ni la "voix passive" en invoquant des raisons de "lisibilité", et assurer une "compréhension facile". Cependant, malgré l'emploi de termes propres au domaine, il ne s'agit pas de recommandations fondées sur des connaissances grammaticales. À titre d'exemple, aucune mention n'est faite des conjonctions relatives dans la recommandation d'éviter les conjonctions, et la voix passive n'apporte pas systématiquement une ambiguïté si les rôles thématiques du verbe sont suffisamment contraignant pour ne pas avoir à se demander si une action doit être réalisée par un opérateur ou une machine (animé/inanimé).

## 2.2 Outils et techniques pour contrôler l'ambiguïté dans les exigences

D'autres outils interviennent après la phase de rédaction, comme *RQA* qui propose une estimation globale de la qualité à partir de mesures de phénomènes de bas niveau, ce qui peut aider le rédacteur à reformuler des passages pendant la correction (Génova *et al.*, 2013). Deux autres approches plus contraignantes que les règles et les guides (désormais confondus sous le terme *recommandations*) de rédaction existent pour uniformiser et faciliter le traitement des spécifications : les modèles à compléter (communément appelés *templates* ou *boilerplates*) et les langues contrôlées. Ces dernières ont fait l'objet de nombreuses études encore d'actualité comme en atteste l'état de l'art approfondi de Kuhn (2014), qui retrace les travaux précurseurs du *Basic English* (Ogden & Graham, 1930) jusqu'à ceux de l'*ASD Simplified Technical English* (ASD, 2013). Ces deux approches (modèles et langues controlées) se positionnent dès les stades d'élaboration et de rédaction des documents, ce qui contraint la liberté du rédacteur qui doit transposer sa représentation de ses connaissances dans un cadre qui supporte peu l'évolution technique et terminologique.

Malgré la diversité de ces approches, le but commun d'optimiser la compréhension des lecteurs - qui seront amenés à réaliser, tester ou utiliser le système - est fondamental pour limiter la prise de risques majeurs, qu'ils soient humains, environnementaux ou financiers. L'étude de phénomènes syntaxiques, sémantiques et lexicaux pour identifier l'ambiguïté doit permettre de constituer des lexiques de termes spécifiques au domaine qui serviront de base de connaissances pour des traitements automatiques divers.

La section suivante présente notre corpus, qui se compose d'exigences fonctionnelles (spécifications techniques de conception de moteurs d'aéronefs).

# 3 Exigences fonctionnelles pour la conception d'un système : description du corpus

Dans cette section, nous présentons les principales caractéristiques qualitatives et quantitatives de notre corpus, ainsi que des extraits illustrant les types d'ambiguïté qui s'y trouvent.

Le corpus que nous explorons contient 5 spécifications d'une société industrielle spécialisée en

conception de moteurs. Certaines de ces spécifications sont à un niveau de rédaction très avancé, voire mature, c'est-à-dire que le document a été validé ou pourrait l'être en l'état par une personne agréée pour être utilisé en phase de développement. Les autres spécifications sont encore à un stade où il manque des informations techniques, ce qui forme des attentes inachevées (112 occurrences de "TBC" pour *To Be Confirmed* et 82 occurrences de "TBD" pour *To Be Determined*) et certaines sections ne sont pas remplies. Il y a 5 186 exigences balisées, mais les 5 594 occurrences de "shall", modal privilégié des exigences d'après toutes les recommandations dédiées à la rédaction en anglais, indiquent la présence d'exigences complexes, comme les exemples suivants en attestent :

*"The electrical actuator **shall** be irreversible : the butterfly **shall** stay in the latest commanded position in the absence of power supply."*
*"XX controller **shall** have 2 independent and dissimilar channels (A & B). Channel A **shall** be a single digital channel and Channel B a segregated full hardware alarm monitoring channel."*

Après concaténation et conversion des documents du format *Word* vers un format de texte brut, les figures qui accompagnent les descriptions textuelles ont été perdues, mais nous conservons les sommaires, titres de sections et sous-sections ainsi que les tableaux. Ces spécifications sont rédigées intégralement en anglais bien qu'il ne s'agisse pas de la langue maternelle de la majorité des rédacteurs [5], et elles représentent plus de 220 000 tokens pour environ 4 500 types. Ce ratio type/token indique qu'il s'agit de données présentant une diversité lexicale très faible, et une rapide observation du corpus démontre une redondance syntaxique très forte. Des échanges avec des rédacteurs et correcteurs de spécifications ont souligné le recours très habituel aux fonctions *copier* et *coller*, notamment pour gagner du temps quand un seul élément change d'une exigence à l'autre. Cependant, cette habitude est également à l'origine d'erreurs redondantes, comme des fautes de frappe ou des fautes d'accord.

La spécificité qualitative de notre corpus repose à la fois sur son domaine (conception de moteurs d'aéronefs) et son degré de spécialisation technique (à l'échelle d'une entreprise), ce qui limite la quantité de données qu'il contient. Nous cherchons donc des marqueurs linguistiques stables et représentatifs de ce corpus pour l'étendre et appliquer des traitements statistiques solides pour en étudier les relations paradigmatiques (des classes de termes flous et de termes génériques/spécifiques).

# 4 Principales sources d'ambiguïté dans les exigences

La notion d'ambiguïté inhérente au langage humain est un phénomène commun qui peut être décrit selon des axes syntaxiques, sémantiques et lexicaux (Zhang, 1998). Cette section s'inspire des travaux de Tjong (2008) et Warnier (2015) ainsi que des cas rencontrés en corpus pour présenter une partie de l'ambiguïté telle qu'elle se manifeste dans les exigences. Nous distinguons l'ambiguïté des exigences (intra-exigences) et l'ambiguïté des spécifications (ambiguïté entre le titre et le contenu de la section, ou ambiguïté dans les renvois à des tableaux ou des figures par exemple) pour ne traiter que l'ambiguïté des exigences à ce stade. Les cas présentés sont associés à une recommandation de rédaction et au nombre de cas rencontrés en corpus pour illustrer l'ampleur de chaque phénomène dans ce corpus. Ils sont organisés en fonction des moyens requis pour automatiser leur détection (lexiques isolés, lexiques associés à des règles symbolique et/ou traitements linguistiques plus profonds).

5. De nombreux indices laissent penser que la langue française interfère, tels que "mesurement" au lieu de "measurement", "discretes inputs" au lieu de "discrete inputs", etc.

## 4.1 Ambiguïté détectable grâce à des lexiques stables

Dans cette section, nous présentons trois types d'ambiguïté qui peuvent être détectés automatiquement grâce à des lexiques qui nécessitent d'être créés mais pas complétés par la suite puisqu'il s'agit simplement de permettre un accès à des classes fermées (grammaticales ou lexicales).

### 4.1.1 Quantifieurs flous

La nécessité de pouvoir mesurer une exigence fait partie des recommandations de rédaction les plus communes. Cela proscrit l'emploi de quantifieurs flous avec des valeurs numériques, pourtant leur présence en corpus montre qu'il s'agit d'un usage qui subsiste. Dans l'exemple suivant, "about" indique que la valeur n'est pas nécessairement exactement "150C" mais ne précise pas les intervalles autorisés :
"*DELETED Even in XX mode, with bleed temperature* ***about 150C****, the ozone concentration shall respect the certification requirements for cabin ozone concentration.*

Le cas suivant est encore plus problématique puisque le sujet indique qu'il s'agit du "maximum", ce qui est contradictoire avec l'attente d'un intervalle autorisé :
"*The* ***maximum*** *confirmed leak/overheat detection time delay shall be* ***around 500 ms****"*

Un simple lexique répertoriant les termes qui amènent une interprétation floue au contact d'une valeur numérique ("about", "approximately", etc.) permet d'automatiser la détection de ce type d'ambiguïté à moindre coût puisqu'il s'agit d'une classe fermée. Nous en détectons 2 répertoriés dans le lexique de termes flous.

### 4.1.2 Références anaphoriques incertaines

Les guides de rédaction des spécifications préconisent de ne pas utiliser de pronoms afin d'éviter les références incertaines. Pourtant, ces références incertaines, qu'elles se rencontrent sous la forme de pronoms ou d'adjectifs possessifs, ne sont pas rares, comme l'illustrent les extraits suivants.

Cet exemple est le cas typique d'une reprise anaphorique incertaine puisque "it" peut renvoyer à "target" ou "altitude" :
"*The flow control target is computed on the A/C altitude based on the ventilation, pressurization and cooling needs.* ***It*** *is dependant of the number of XX in control."*

D'autres cas problématiques apparaissent aussi quand le pronom fait référence à un groupe nominal antérieur mais que celui-ci n'est pas présent dans l'exigence (il peut être explicite dans l'exigence précédent, mais elles sont censées être autonomes et complètes selon tous les standards) :
"***It*** *shall be selectable by the pilot without any restriction."*

Dans l'exemple qui suit, l'adjectif possessif "their" ne peut se remplacer que par une paraphrase qui complexifierait sensiblement la phrase d'origine. De plus, l'usage du pluriel peut prêter à confusion quant au lien entre chacun des deux "XX P." et les "associated XX'. Ainsi :
"*Both XX P. shall be ISOLATED by commanding closed* ***their*** *associated XX in case of DITCHING mode activation."*

pourrait devenir :
*Both XX P. shall be ISOLATED by commanding closed the XX associated with* ***both/each*** *XX P. in case of DITCHING mode activation.*

Encore une fois, il s'agit de termes issus de classes restreintes qui peuvent être organisé sous forme de lexiques pour détecter leur présence. Si une certaine tolérance est accordée (vérification des références

possibles en fonction du genre et du nombre des groupes nominaux déjà rencontrés dans la même exigence), une combinaison de ces lexiques à des patrons morpho-syntaxiques peut assouplir la détection de références incertaines. Sans considérer les contextes d'apparition, 40 occurrences de la forme "it" sont automatiquement repérées en corpus, ainsi que 3 occurrences de "their".

### 4.1.3 Portée des conjonctions (and, or, either, ...) et de la négation

Les guides de rédaction mettent en garde contre l'emploi de conjonctions qui complexifient l'exigence. En effet, l'emploi de conjonctions peut amener des doutes quant à leur portée, voire de leur priorité quand plusieurs sont combinées.

L'exemple suivant est un cas critique et si la présence des conjonctions complexifie le message, la longueur de la phrase est un aspect trivial qui y contribue également :
*"The XX shall automatically control the system operation **and** configuration for continued safe flight **and** landing in the event of a failure within the system **or** interfacing system, **except for** emergency ram air selection **and** auxiliary pressurization selection."*

L'expression d'exigences négatives (capacités non-attendues, actions à ne pas effectuer, etc.) sont également signalées comme indésirables par les guides de rédaction. En effet, s'il est parfois pertinent de préciser ce qu'un système ne doit pas faire, d'un point de vue strictement linguistique la portée d'une négation peut être source d'ambiguïtés, notamment en présence de conjonctions comme ici :
*"The XX shall **not** suffer any deformation **and** operates*[6] *normally after the application of clamp assembly torque 100% higher than the nominal torque defined by XX in the XX."*

La décomposition de la phrase résulte en différentes transformations, sans compter la présence du "s" final du second verbe, preuve de la multiplicité d'interprétations potentielles. La reprise du sujet après la conjonction de coordination facilite la mise en parallèle des deux instructions :
*The XX shall **not** suffer any deformation **and** the XX shall (**not ?**) operate<u>s</u> normally [...]*

La détection des adverbes de négation ("not", "never", "no... more", etc.) est une tâche automatisable simplement par l'accès à des lexiques spécifiques. 303 conjonctions de coordination sont détectées comme problématiques dans le corpus, parfois associées à l'une des 226 occurrences signalées de l'adverbe de négation "not".

## 4.2 Ambiguïté détectable grâce à des lexiques spécifiques et des patrons morpho-syntaxiques

Les exemples présentés jusqu'ici nécessitent d'accéder à un lexique fini pour automatiser la détection, ce qui est une tâche relativement simple à partir du moment où les éléments contenus dans une classe fermée sont tous connus. Les cas présentés par la suite requièrent un traitement plus complexe, que ça soit pour la détection de certains phénomènes ou pour assurer la complétion semi-automatique des ressources nécessaires.

### 4.2.1 Rattachement prépositionnel

Les phrases longues et complexes sont décriées par les guides de recommandations, qui conseillent d'être "concis" et de rédiger des phrases "simples". Malgré une absence de précision concernant le nombre de mots à ne pas dépasser, il est clair qu'une phrase longue est un terrain idéal pour éloigner

6. Les extraits présentés ne sont pas corrigés

un élément de son ou ses complément-s, et la distance entre un syntagme et un complément peut servir d'indication stable pour évaluer la complexité d'une phrase (Tanguy & Tulechki, 2009).

Dans l'exemple suivant, le syntagme prépositionnel "on its safety channel and on one main channel" peut se rapporter au noyau verbal "receive" ou au noyau nominal "element". Sans précision supplémentaire c'est uniquement une connaissance avancée qui sélectionne l'interprétation attendue :
*"Each XX Controller shall **receive** the signal from one sensing **element** of each Inbd WAITS dual sensing element skin temperature sensor **on its safety channel and on one main channel**."*

Les étiqueteurs syntaxiques sont encore facilement perturbés par des données trop différentes de celles pour lesquelles ils ont été entraînés et parmi ceux librement distribués, peu d'entre eux sont performants sur des exigences sans adaptation préalable. Cependant, en combinant des informations morpho-syntaxiques sur les phrases et des connaissances accessibles automatiquement sur les rôles thématiques des verbes rencontrés (nature et nombre d'arguments obligatoires et facultatifs) la détection automatique de ces relations à longue distance est envisageable. Ce phénomène n'est pas encore détecté automatiquement par *Semios* mais est envisagé comme un axe d'amélioration pour assurer la qualité des exigences.

### 4.2.2 Voix passive et absence d'agent

Les recommandations de rédaction ne sont pas systématiquement justifiées explicitement et c'est le cas concernant l'usage de la voix passive. L'exemple qui accompagne le conseil de recourir uniquement à la voix active est celui-ci *"shall be able to select"*. Il est difficile d'affirmer si c'est le risque de ne pas préciser qui ou quoi doit effectuer l'action qui est vu comme un problème, ou la complexité évitable de la phrase induite par cette structure. Dans le corpus exploré, la séquence "shall be able to" est présente 19 fois, dont 4 fois "shall be able to be". Plus généralement, le corpus compte un très grand nombre de phrases à la voix passive (plus de 7 000 occurrences de "be + participe passé").

Dans les extraits suivants, l'agent qui doit effectuer l'action n'est jamais désigné, par exemple de façon à distinguer les cas où l'action doit être effectuée par une machine ou par un agent humain :
*"The leak shall **be detected** when the insulation resistance is lower than 150 ohms."*
*"The CPCS cabin altitude limitation shall **be triggered** when the cabin altitude (ZC) is superior of equal to the SAFETY_ALT_LIM setting."*

L'automatisation de la détection de la voix passive par des étiqueteurs donne également des résultats peu fiables, mais le recours à des informations morpho-syntaxiques de surface permet de détecter une partie des cas présents en corpus et de filtrer les phrases dans lesquelles les agents sont désignés explicitement par la préposition "by". De cette façon, 944 cas sont identifiés. En ce qui concerne les structures complexes, 14 sont détectées, toujours grâce à la combinaison de patrons syntaxiques et de lexiques spécifiques.

### 4.2.3 Énumérations

Les énumérations sont très présentes dans les exigences mais il est courant qu'elles ne contiennent pas d'information explicite concernant les relations qu'entretiennent leurs éléments. C'est le cas de l'exemple suivant qui ne précise pas si une des conditions suffit ou si elles doivent toutes être réunies, et qui présente un problème de niveaux d'arborescence dans le document original :

"LP leak for isolation shall be computed every 100 ms and :
— set if the following conditions are gathered :

— LP leak is detected
— Right engine start is not requested
— reset if following conditions are gathered :
— LP leak is not detected
— Left bleed P/B is OFF"

Du point de vue du traitement automatique, une fois la variabilité de présentation (listes à puce, numérotées, indentation, ponctuation ou non, etc.) prise en compte, il devient possible de détecter les énumérations incomplètes. En outre, si l'attribution des parties du discours est correctement effectuée (stade important et encore très dépendant de l'outil d'étiquetage choisi), il devient également possible de révéler celles qui ne coordonnent pas des syntagmes équivalents comme dans l'exemple précédent qui ne respecte pas les différents niveaux en alternant des actions d'action ("set" et "reset") et des états ("LP leak is detected"). Le traitement de ce phénomène n'est pas encore exploitable par l'outil présenté ici.

### 4.2.4 Polysémie, homonymie et polyacception

Des guides de rédaction recommandent qu'un terme utilisé en tant que nom ne soit pas utilisé en tant que verbe (par exemple "set" dans la langue générale). Une autre restriction lexicale concerne la pluralité des sens d'un terme, soit la polysémie si ses différents sens sont liés ou l'homonymie si ses sens ne sont pas liés. A priori, mêmes confondus, ces phénomènes sont peu rencontré dans nos données, cependant, il reste possible d'en trouver quelques exemples en corpus, à l'instar de la forme "current" qui apparaît dans le corpus alternativement en tant que nom et adjectif :

*"The maximum **current** per phase has also to be specified"*
*"28VDC state **current** shall be able to supply output with 400mA Nominal"*

*"XX shall be designed to stay in the **current** position in case of power supply loss up to the maximum possible cabin airflow conditions."*
*"**Current** XX shall only be erased when data loading is correct"*

Cela peut amener une confusion quand le terme est placé immédiatement à gauche d'un nom (rivalité entre la position canonique des adjectifs épithètes et la construction usuelle des syntagmes Nom-Nom en anglais), notamment pour les locuteurs dont l'anglais n'est pas la langue maternelle :
*"Over **current** protection circuit shall be activated at min 500 mA whatever the duration, and at 6.5A max to protect the XX hardware."*

Pour ce type d'ambiguïtés, différents niveaux d'analyse (morpho-syntaxiques et distributionnels) doivent être combinés pour distinguer automatiquement la nature du terme et évaluer si plusieurs acceptions ("polyacception" chez Condamines & Rebeyrolle (1996)) cohabitent en fonction des contextes. 617 occurrences de 70 adjectifs et adverbes flous contenus dans un lexique construit et complété manuellement sont détectés. Le recours à une méthode semi-automatique pour compléter de tels lexiques permettrait de les enrichir, potentiellement pour distinguer si l'ambiguïté provient d'une pluralité de sens ou de natures grammaticales. Une telle précision dans l'analyse pourrait même guider implicitement le rédacteur vers une substitution ou reformulation adaptée.

### 4.2.5 Sous-spécification

Les guides de rédaction mettent au premier plan la nécessité de disposer d'exigences complètes et la sous-spécification ("generality" chez (Zhang, 1998)) en est l'une des manifestations possibles. Elle peut se retrouver au niveau des constituants de la phrase, nominaux comme verbaux. Elle se traduit

par l'absence de compléments syntaxiquement facultatifs mais sémantiquement distinctifs.

Dans l'exemple suivant, il est quasiment impossible de savoir à quel système le rédacteur fait référence étant donné qu'il s'agit d'un cas typique de terme générique. En effet, "system" n'est pas accompagné d'un complément, qui pourrait être un nom (par exemple "pressurization system") ou une relative ("The system, which monitors..." ou "The system that maintains..."), et il semble peu probable qu'un aéronef comporte un système unique.

*"When the aircraft leaves the ground, the takeoff sequence is initiated.* ***The system*** *will maintain the pre-pressurization cabin altitude of a set amount below field elevation until the aircraft reaches a set altitude or differential from takeoff, or a set time has elapsed."*

À l'inverse, *"aircraft"* peut se passer de complément, à condition qu'il s'agisse du thème central de la spécification qui serait cité d'une exigence à l'autre.

Ce cas, le plus complexe que nous ayons rencontré en corpus autant d'un point de vue linguistique que de l'automatisation de sa détection, se distingue de la polysémie puisque *"system"* a toujours la même signification, mais c'est sa généricité qui le rend difficilement interprétable d'une seule façon. Le degré de généricité étant, au moins partiellement, dépendant du domaine, sans lexiques de termes spécifiques au domaine, la détection de termes incomplets ne peut qu'échouer. L'hypothèse qui sera testée suite à cette étude avance que la complétion semi-automatique de ressources contenant les termes polylexicaux (issus de relations syntagmatiques privilégiées) est envisageable en utilisant les comportements en contexte des termes déjà connus comme des amorces.

## Bilan de la détection automatique de l'ambiguïté dans les spécifications

Ce panorama d'exigences choisies met en avant la multiplicité des risques liés à l'ambiguïté d'un message technique destiné à servir de support pour la conception de systèmes et en cas de situation anormale. Ces risques peuvent être critiques, comme dans les cas suivants : 1) l'utilisateur, qu'il voie ou non les différentes interprétations possibles, peut opter pour l'une d'entre elles qui ne respecte pas les attentes du rédacteur, ou 2) l'utilisateur peut détecter différentes interprétations et ne pas savoir laquelle préférer, ce qui ralentit la mise en exécution de la suite de la conception. Un autre type de risque est de se confronter à un échec du projet sans en avoir identifié la ou les cause-s, ce qui constituerait une cause d'échec potentielle pour un projet ultérieur.

Tous les cas d'ambiguité ne se situent pas au même niveau (syntaxique, sémantique ou lexical) et pour cette raison ne se détectent pas de la même façon. Ainsi, l'ambiguïté portée par les quantifieurs flous, les références anaphoriques, la portée des conjonctions de coordination et de la négation est identifiable par des lexiques génériques. À l'inverse, des lexiques dépendants du domaine et du cadre du processus de conception sont nécessaires pour détecter les ambiguïtés émanant des cas de polysémie, homonymie, polyacception et sous-spécification. Quant à la détection des ambiguïtés liées au rattachement prépositionnel, à la voix passive et aux énumérations, elle nécessite des analyses morpho-syntaxiques, parfois combinées à différents types de lexiques.

Dans cette démarche, Prometil commercialise *Semios*, une branche dédiée à l'analyse automatique des exigences rédigées en langage naturel issue d'un outil développé par l'Institut de Recherche en Informatique de Toulouse (IRIT) et Prometil, *Lelie* (Kang & Saint-Dizier, 2015). *Semios* fonctionne grâce au moteur *TextCoop* autour duquel s'articulent des lexiques et des patron morpho-syntaxiques dans un format *Prolog* qui sont adaptés manuellement, après une exploration approfondie de nouvelles spécifications, lorsqu'un nouveau domaine doit être analysé.

Les lexiques de cet outil contiennent différents niveaux d'informations, comme le degré de flou attribué aux adjectifs ("normal" est moins flou que "high" selon le barème interne) ou des informations morphologiques (ce qui permet de pallier partiellement l'absence d'un étiqueteur morpho-syntaxique). Les lexiques sont également interrogés lors de l'analyse de certains patrons morpho-syntaxiques, comme ceux qui détectent la combinaison du modal "be" suivi (directement ou non) d'un participe passé ou les structures verbales complexes. Une série de règles symboliques détectent l'emploi de pronoms (contenus dans un lexique) dont la référence n'est pas explicite ou qui n'ont pas du tout d'antécédents. Le tableau 1 présente une synthèse des cas d'ambiguïtés automatiquement détectés et associés à des recommandations issues de différents standards de rédaction. Certains traitements répondent à plus d'une recommandation mais nous indiquons seulement l'une d'elles le cas échéant.

L'objectif final du contrôle de l'ambiguïté, selon notre approche, n'est pas d'imputer des responsabilités aux acteurs impliqués dans le processus de conception mais de limiter les échecs de réalisation de projets. L'analyse du contenu en phase de relecture offre un avantage non-négligeable comparé aux autres outils et techniques à disposition des rédacteurs de spécifications techniques qui se placent majoritairement bien plus tôt dans l'élaboration des documents (langages contrôlés ou templates). Le contenu exprimé dans un cadre relativement libre (les variations lexicales et syntaxiques de nos spécifications sont plus pauvres que dans des corpus de langue générale, ce qui s'observe par exemple à travers le ratio type/token présenté dans la section 3) permet au rédacteur de respecter au mieux sa représentation des concepts et connnaissances du domaine. Sans aborder la question de la limitation de l'erreur humaine par cette approche, la relecture automatique analyse un contenu représentatif de la réalité du domaine, ce qui peut alimenter la construction de ressources linguistiques pertinentes, c'est-à-dire structurées et spécifiques.

| **Recommandations** | **Détections associées** | **Nombre de cas** |
|---|---|---|
| Références claires | Voix passive (be + p.passé) sans "by" | 944 |
| | Termes flous (normal, low, more, ...) | 617 |
| | Pronoms (it, their) | 43 |
| Phrases simples | Conjonctions de coordination (and, or) | 303 |
| | Adverbe de négation (not) | 226 |
| | Structures verbales complexes (be able to, ...) | 14 |
| Exigence mesurable | Quantifieurs flous (about, around + valeur num.) | 2 |
| | | **Total : 2 149** |

TABLE 1: Ambiguïté détectée automatiquement (n = 5 186 exigences)

Compte tenu du nombre théorique d'exigences (5 186 identifiants uniques) et du nombre d'alertes (2 149), le résultat lissé indique qu'une exigence sur deux contient une ambiguïté potentielle. Comme l'ont illustré une grande partie des exemples, il est commun que plusieurs phénomènes liés à l'ambiguïté se combinent au sein d'une même exigence.

Étant donné la quantité de travaux menés à propos de l'ambiguïté véhiculée par les éléments grammaticaux (connecteurs, marques de la négation et du pluriel, etc.), nous privilégions l'investigation de l'ambiguïté lexicale et sémantique, typiquement les cas de polysémie, homonymie, homographie et sous-spécification nominales. Nous envisageons en outre de constituer un lexique de verbes à partir d'une étude de leurs arguments obligatoires et facultatifs ainsi que de leurs rôles thématiques pour traiter la sous-spécification verbale. La détection de ces manifestations fait encore partie des tâches difficiles à automatiser dans les données qui disposent de peu de ressources adaptées. Les premiers cas identifiés manuellement doivent permettre de détecter automatiquement des éléments lexicaux

présentant des comportements équivalents en contexte (axes syntagmatique et paradigmatique). La section qui suit expose les prochaines étapes pour automatiser la détection de termes ambigus.

# 5 Perspectives

Dans cet article, nous présentons la première étape de mise au point d'une méthodologie robuste permettant de constituer semi-automatiquement des ressources lexicales adaptées aux domaines techniques spécialisés. Cette étape consiste à caractériser les phénomènes d'ambiguïté syntaxique, sémantique et lexicale en vue d'adapter les approches requises pour les détecter automatiquement et enrichir des ressources en limitant toute intervention humaine. À partir de l'hypothèse Harrisienne qui postule que des termes partageant des contextes (i.e. des voisins distributionnels) présentent une relation de synonymie, d'hyperonymie ou un lien sémantique plus lâche (Harris, 1954), nous envisageons l'analyse distributionnelle automatique (ADA) comme une piste pertinente pour construire des ressources sémantiques spécifiques. Cependant, reposant sur des principes de fréquences relatives et des répartitions statistiques, l'ADA nécessite d'être appliquée sur des données suffisamment massives, à l'instar des corpus construits ces dernières années à partir de la richesse du web (Tanguy, 2012), pour conclure à des analyses pertinentes vis-à-vis du cadre linguistique étudié.

Si des adaptations de l'analyse distributionnelle automatique à des corpus de petite taille ont abouti (Fabre *et al.*, 2014), le recours a des corpus plus riches (en nombre de mots ainsi qu'en variations lexicales et syntaxiques) permet de limiter l'intervention humaine pour interpréter des résultats trop disparates. Parmi les techniques d'expansion semi-automatique de corpus, la chaîne de traitement *BootCaT* (Baroni & Bernardini, 2004) consiste à aspirer des pages web grâce à des requêtes itératives par mots-clefs complexes (*tuples*). Si cette technique permet de constituer facilement un corpus de pages web contenant les tuples, la mesure de la similarité entre les documents obtenus et le corpus de départ reste un aspect qui demande d'être étudié en détails, notamment grâce à la grille de dimensions de Sinclair (1996). Selon l'étude de Warnier (2015), l'observation de phénomènes tels que la fréquence relative de parties des pronoms et négations s'avère utile pour gagner en précision dans la caractérisation du genre et du registre. Cette finesse de description semble cruciale pour appréhender les caractéristiques d'un corpus acquis de manière semi-supervisée et mesurer l'impact du choix des tuples. À titre d'exemple, les extraits suivants sont issus du corpus que nous avons construit après sélection de 51 tuples via une extraction terminologique avec YaTeA (Hamon, 2012) et des calculs de fréquence fournis par le concordancier AntConc (Anthony, 2005).

*"Aerocon worked closely with the Federal Aviation Administration to establish the safety standards for the Lower Lobe Cargo Compartment, the first of its kind combining a large occupant complement, full-featured lavatories, and Class B stowage, and is the only company in the world offering this system with STC, configured to your requirements."* [7]

*"A flight attendant has the same benefits as other airline employees, such as paid vacation, paid sick leave, paid medical insurance and life insurance, and retirement benefits, greatly reduced air travel expenses for self and immediate family, and credit union membership. "* [8]

*"If you develop gallstones, for example, the traditional treatment has long been cholecystectomy (gall bladder removal), which is major abdominal surgery in which the surgeon removes the gall bladder through a 5- to 8-inch incision. Recovery typically involves a week of post-surgical hospitalization, followed by several weeks of recovery at home. This standard treatment is extremely invasive, and so*

7. http ://aeroconengineering.com/crc.html
8. http ://avstop.com/careers/flightattendants.html

*not surprisingly the incidence of complications and even death is significant. (My dad very nearly died as the result of complications following an open cholecystectomy operation).*" [9]

Ce dernier exemple met également l'accent sur des facteurs à contrôler pour obtenir le corpus le plus proche possible du corpus d'exigences : la proportion de contenu technique par rapport au contenu connexe au thème principal de ces exigences (conception de moteurs). Dans le corpus-web, les pages abordant la santé des techniciens issus de l'aéronautique sont communes, notamment sur les forums, tandis qu'il s'agit d'un sujet qui n'est jamais abordé dans les spécifications. Un exemple moins extrême et plus courant concerne les références aux textes relatifs au trafic aérien qui constitue un thème à part entière dans le corpus-web, contrairement à sa place dans notre corpus où il est cité mais jamais développé.

Si la comparaison ne permet pas de conclure à une similarité parfaite entre les deux corpus, ils partagent toutefois des traits assez proches (multiplicité de rédacteurs, unicité du thème, densité relative du vocabulaire, effets visés). En appliquant les traitements mis en œuvre pour sélectionner les tuples identifiés comme spécifiques au corpus d'exigences (extraction terminologique, observation des fréquences relatives et absolues, extraction de patrons morpho-syntaxiques) au corpus-web, nous espérons affiner la récupération des pages web grâce à des tuples plus spécifiques.

L'effet *boîte noire* de la solution adoptée pourrait être amoindri en grâce à une technique alternative, *Synopsis*, qui intègre à sa chaîne de traitement la caractérisation des données aspirées par confrontation d'un corpus contenant les termes complexes des requêtes, *germs*, avec un corpus de la même envergure ne contenant aucun de ces termes (Duthil *et al.*, 2011). Une approche complémentaire serait de généraliser les comportements distributionnels pour calculer la similarité sur des ensembles plus grands (Périnet & Hamon, 2014).

En obtenant une variété de contextes suffisant pour appliquer l'ADA aux corpus techniques spécialisés, nous disposerons du matériau linguistique pour détecter automatiquement le voisinage de termes ambigus (flous ou sous-spécifiés) et enrichir semi-automatiquement des ressources spécialisées.

# 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*.

CONDAMINES A. & REBEYROLLE J. (1996). Point de vue en langue spécialisée. *Meta*, **42**(1), 174–184.

DUTHIL B., TROUSSET F., ROCHE M., DRAY G., PLANTIÉ M., MONTMAIN J. & PONCELET P. (2011). Towards an automatic characterization of criteria. In *Database and Expert Systems Applications*, p. 457–465 : Springer.


9. http ://www.avweb.com/news/savvyaviator/savvy_aviator_53_dark_side_of_maintenance_196909-1.html

FABRE C., HATHOUT N., SAJOUS F. & TANGUY L. (2014). Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille. In *21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, p. 266–279.

GÉNOVA G., FUENTES J. M., LLORENS J., HURTADO O. & MORENO V. (2013). A framework to measure and improve the quality of textual requirements. *Requirements engineering*, **18**(1), 25–41.

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.

IEEE (1998). *IEEE 1220-1994*. Rapport interne, Rapport technique.

IEEE/ISO/IEC (2011). *IEEE, ISO, IEC. 29148 : 2011-Systems and software engineering-Requirements engineering*. Rapport interne, Rapport technique.

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

KUHN T. (2014). A survey and classification of controlled natural languages. *Computational Linguistics*, **40**(1), 121–170.

OGDEN C. K. & GRAHAM E. (1930). *Basic English*. K. Paul.

PÉRINET A. & HAMON T. (2014). Réduction de la dispersion des données par généralisation des contextes distributionnels : application aux textes de spécialité. *Articles longs*, p. 232.

SINCLAIR J. M. (1996). *Preliminary recommendations on text typology*. Birmingham Corpus Linguistics Group School.

TANGUY L. (2012). *Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes*. PhD thesis, Université Toulouse le Mirail-Toulouse II.

TANGUY L. & TULECHKI N. (2009). Sentence complexity in french : A corpus-based approach. *Proceedings of IIS (Recent Advances in Intelligent Information Systems)*, p. 131–145.

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

WARNIER M. (2015). How can corpus linguistics help improve requirements writing ? specifications of a space project as a case study. In *Requirements Engineering Conference (RE), 2015 IEEE 23rd International*, p. 388–392 : IEEE.

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