Analyse d'un corpus d'exigences pour améliorer la rédaction des spécifications de systèmes spatiaux au CNES

Maxime Warnier, Anne Condamines

Analyse d'un corpus d'exigences pour améliorer la rédaction des spécifications de systèmes spatiaux au CNES

Maxime Warnier[1,2], Anne Condamines[1]

[1]CLLE-ERSS (UMR 5263 : CNRS & Université Toulouse – Jean Jaurès) – France

[2]Centre National d'Études Spatiales – France

Abstract

The aim of this work is to improve the clarity and precision of the technical specifications written by the engineers at CNES (Centre National d'Études Spatiales / National Centre for Space Studies) prior to the realization of space systems. The importance of specifications (and particularly of the requirements that are part of them) for the success of large-scale projects is indeed widely acknowledged; similarly, the main risks associated with the use of natural language (ambiguity, vagueness, incompleteness) are relatively well identified. In this context, we are trying to set up a solution that would be adopted by the engineers at CNES (who are currently not asked to follow specific writing rules): therefore, this solution should be both effective (i.e. it should significantly limit the above-mentioned risks) and not too disruptive (which would make it counterproductive). A Controlled Natural Language (CNL) – i.e. a set of linguistic rules constraining the lexicon, the syntax and the semantics – seems to be an interesting option, provided that it remains close enough to natural language. Unfortunately, the CNLs for technical writing that we considered are not always relevant from a linguistic point of view. We would therefore like to develop our own CNL for requirements writing in French at CNES. Our methodology relies on the hypothesis of the existence of a sublanguage; besides, we make use of Natural Language Processing tools and methods to validate the relevance of the rules on a corpus of genuine requirements written for former projects.

Résumé

L'objectif de notre travail est d'augmenter la clarté et la précision des spécifications techniques rédigées par les ingénieurs du CNES (Centre National d'Études Spatiales) préalablement à la réalisation de systèmes spatiaux. L'importance des spécifications (et en particulier des exigences qui les composent) pour la réussite des projets de grande envergure est en effet désormais très largement reconnue ; de même, les principaux risques liés à l'utilisation de la langue naturelle (ambiguïté, flou, incomplétude) sont relativement bien identifiés. Dans ce contexte, nous nous efforçons de mettre au point une solution qui soit réellement adoptée par les ingénieurs du CNES (qui ne sont actuellement pas tenus de suivre des règles de rédaction) : celle-ci se doit donc d'être à la fois efficace (autrement dit, elle doit limiter sensiblement le risque langagier) et aisée à mettre en place (autrement dit, elle ne doit pas bouleverser trop profondément leurs habitudes de travail, ce qui la rendrait contre-productive). Une langue contrôlée, c'est-à-dire un ensemble de règles linguistiques portant sur le vocabulaire, la syntaxe et la sémantique, nous paraît être une réponse idéale à ce double besoin – pour autant qu'elle reste suffisamment proche de la langue naturelle (et particulièrement de l'usage qui en est fait lors de la rédaction des exigences). Or, les langues contrôlées pour la rédaction technique que nous avons envisagées ne nous semblent pas toujours pertinentes d'un point de vue linguistique. Nous voudrions donc définir notre propre langue contrôlée pour la rédaction des exigences en français au CNES. L'originalité de notre démarche consiste à supposer l'existence d'un sous-langage et à systématiquement vérifier nos hypothèses sur des corpus d'exigences authentiques à l'aide de techniques et d'outils de traitement automatique du langage.

Keywords: corpus ; controlled natural language ; requirements ; technical writing ; sublanguage

Mots-clés : corpus ; langue contrôlée ; exigences ; rédaction technique ; sous-langage

1. Introduction

Cette étude fait suite à une demande émanant de la sous-direction Assurance Qualité du CNES (Centre National d'Études Spatiales), l'agence spatiale française, et visant à améliorer la qualité rédactionnelle des spécifications de systèmes spatiaux qui y sont conçus. Les spécifications techniques sont des collections d'*exigences*, elles-mêmes définies comme étant les représentations de conditions ou de fonctionnalités devant être remplies ou possédées par un système pour satisfaire un document formellement imposé, tel un contrat ou une norme[1] (IEEE, 1990). Parce qu'elles influencent la réussite – ou l'échec – des projets concernés et parce qu'elles revêtent une valeur contractuelle lorsque plusieurs partenaires sont impliqués (tel que c'est de plus en plus souvent le cas), l'importance cruciale des exigences est désormais largement acceptée (Nuseibeh et Easterbrook, 2000), au moins en ce qui concerne les projets de grande envergure (c'est-à-dire mobilisant de nombreuses personnes durant une période de temps longue) ; une discipline leur est même dédiée, l'ingénierie des exigences. Dès lors, de nombreux outils et des bonnes pratiques ont vu le jour pour en améliorer la gestion, la traçabilité et, plus en amont, la rédaction.

En effet, les risques associés à l'utilisation de la langue naturelle (Pace et Rosner, 2010) – parmi lesquels on compte en particulier l'ambiguïté (Kamsties et Peach, 2000), le flou (Zhang, 1998) et l'incomplétude – ainsi que leurs possibles conséquences indésirables (allant des retards et des augmentations de coûts aux accidents, en passant par les litiges entre parties prenantes) sont bien connus des responsables de projets et des ingénieurs expérimentés. Soucieuses de les prévenir, les grandes entreprises ou institutions tendent donc à adopter des solutions visant, sinon à les éliminer, du moins à les limiter à un niveau plus acceptable. Ces solutions préventives sont de natures diverses (langages plus formels (Meyer, 1985) ou outils d'assistance à la rédaction (Carlson et Laplante, 2013), par exemple) et peuvent donc être plus ou moins radicales (les formalismes qui tendent vers l'univocité étant généralement les plus contraignants). À l'heure actuelle, aucune de ces solutions n'est toutefois imposée aux ingénieurs du CNES en charge de la rédaction des exigences, ce qui signifie qu'ils sont libres d'écrire les exigences sans restrictions et de la façon qu'ils jugent intuitivement préférable.

En dépit des inconvénients cités précédemment, la langue naturelle (en l'occurrence, le français) possède également des caractéristiques intéressantes, qui justifient que l'on cherche à les préserver : d'une part, elle permet une expressivité maximale (tout ce qui peut être exprimé dans un quelconque langage peut l'être en langue naturelle, l'inverse n'étant pas vrai) et, d'autre part, elle est particulièrement simple à utiliser (tant pour les rédacteurs que pour les lecteurs), puisqu'elle ne nécessite pas d'être réapprise. Ces deux avantages non négligeables la rendent par ailleurs inévitable à un moment ou à un autre du processus de spécification.

Par conséquent, le compromis que nous souhaitons adopter dans l'optique de proposer une solution à la fois efficace (réduisant le risque langagier) et adaptée (facilement utilisable par les ingénieurs) consiste à mettre au point un ensemble de règles de rédaction qui éliminent les potentiels risques de mauvaise interprétation. Un tel ensemble de règles est connu dans la littérature sous le nom de langue contrôlée (LC) – ou *Controlled (Natural) Language* (CL/CNL) en anglais. Selon Kuhn (2014), un langage peut être appelé langue contrôlée s'il respecte les quatre propriétés suivantes : (1) il est basé sur une seule langue naturelle, (2) il est

[1] "requirement: […] condition or capability that must be met or possessed by a system or a system component to satisfy a contract, standard, specification, or other formally imposed document […]"

plus restrictif en ce qui concerne le lexique, la syntaxe et/ou la sémantique, (3) il reste (au moins substantiellement) compréhensible par les locuteurs de la langue source et (4) il s'agit d'un langage consciemment défini. Le langage contrôlé que nous souhaitons élaborer se doit en outre de rester aussi proche que possible des habitudes des rédacteurs (autrement dit, de ce que l'on trouve déjà dans des spécifications précédentes) ; en effet, si les règles qui le constituent sont ressenties comme une contrainte trop forte parce qu'elles sont trop éloignées de la réalité de la pratique – comme c'est selon nous le cas des règles trouvées dans d'autres langues contrôlées existantes –, elles risquent au mieux d'être ignorées, au pire de se révéler contre-productives. Pour ce faire, nous faisons l'hypothèse de l'existence d'un sous-langage, qui pourrait servir de base pour de nouvelles règles.

2. Sous-langage et langue contrôlée

Comme précisé en introduction, nous voudrions que les règles qui constituent notre langue contrôlée pour la rédaction des exigences soient en adéquation avec la pratique effective des ingénieurs du CNES. À cette fin, nous cherchons donc dans un premier temps à identifier d'éventuelles régularités langagières dans les productions de ces derniers, lesquelles nous serviraient de base concrète pour l'élaboration des règles de rédaction.

Il peut *a priori* sembler contradictoire de chercher des régularités dans des écrits qui, comme déjà dit, n'ont justement pas été rédigés en suivant des règles imposées – ce qui revient à dire qu'il pourrait très bien y avoir autant de styles de rédactions que de rédacteurs différents au CNES. Nous faisons néanmoins l'hypothèse que puisse exister un sous-langage propre aux exigences[2]. Un *sous-langage* est défini comme un sous-ensemble de la langue générale, émergeant spontanément dans des domaines sémantiques restreints, utilisé par une communauté de spécialistes dans des situations récurrentes[3] et divergeant de la langue générale notamment par (1) une syntaxe, une structure textuelle et un lexique restreints, (2) une syntaxe dite « déviante » et (3) des fréquences différentes d'occurrences de mots et de motifs syntaxiques (définition synthétisée par Temnikova et al., 2014). Les conditions d'apparition d'un sous-langage étant ici clairement rencontrées (communauté d'experts, domaine restreint, récurrence des situations), nous pouvons nous attendre à en repérer les manifestations (lexicales et morphosyntaxiques en particulier) au sein des exigences constituant notre corpus. D'autres facteurs peuvent également être à l'origine de régularités dans les exigences, notamment l'influence de règles de rédaction dont les ingénieurs ont pu prendre connaissance par ailleurs (même si elles ne leur sont pas imposées), les révisions effectuées par les chefs de projets avant que les spécifications ne soient définitivement approuvées ou simplement des intuitions partagées à propos de certains phénomènes langagiers.

L'objectif de notre démarche est donc d'élaborer une langue contrôlée[4] (langage plus restreint que la langue naturelle, défini consciemment) basée sur un sous-langage (sous-ensemble de la

[2] Pour être exact, à ce stade, nous nous intéressons au sous-langage des exigences *en français au CNES*.

[3] Notons qu'ainsi définie, cette notion est très proche de celle de *genre textuel* (cf. entre autres la définition de *textual genre* par Bhatia (1993), d'après le travail de Swales : "recognizable communicative event characterized by a set of communicative purpose(s) identified and mutually understood by the members of the professional or academic community in which it regularly occurs").

[4] Que l'on peut qualifier de « naturaliste » (Clark et al., 2010).

langue générale, apparaissant spontanément)[5] ; la première incarne le versant prescriptif (parce qu'elle vise à réguler l'emploi de la langue pour en éliminer les structures potentiellement problématiques), là où le second représente le versant plus descriptif (parce qu'il consiste à recenser les structures les plus fréquentes) d'une même pratique, celle de la rédaction des exigences. En identifiant l'ensemble relativement clos de structures syntaxiques et le vocabulaire du sous-langage[6] (Somers, 1998), nous nous efforçons de rester proches des habitudes de rédaction des ingénieurs ; en imposant ou interdisant certaines structures linguistiques, nous essayons de limiter le risque langagier. Par ailleurs, ces deux versants peuvent aisément être rapprochés des notions de *normalisation* et *normaison*, respectivement : Gaudin (1993) précise que « l'analyse tirerait profit à opposer deux procès normatifs : la normaison, relevant de l'activité spontanée à l'œuvre dans tout échange, et la normalisation, domaine des interventions conscientes et planifiées » (voir aussi Condamines et Warnier (à paraître)).

3. Corpus

Pour pouvoir vérifier notre hypothèse initiale, nous avons constitué un corpus d'exigences en français, extraites automatiquement de spécifications fournies par le CNES. Celles-ci concernent deux projets spatiaux différents, appelés respectivement Pléiades (15 spécifications, 2 500 exigences, 120 000 mots environ) et Microscope (15 spécifications, 1 000 exigences, 44 000 mots environ). Ces spécifications représentent différents niveaux de l'« arbre produit » des deux projets (système, sous-système, interface) et ont été écrites par des rédacteurs au profil similaire (généralement issus d'une école d'ingénieurs française, puis formés en interne). Les deux projets se distinguent cependant par leur ampleur (Pléiades étant nettement plus important que Microscope, ce qui se traduit par davantage d'exigences) et par leur nature et domaine (Pléiades consiste en deux satellites d'observation haute résolution de la Terre, tandis que l'objectif de Microscope est de réaliser une expérience scientifique validant le principe physique d'équivalence faible).

Quoique de tailles modestes (ce qui est souvent le cas dans les domaines spécialisés et s'explique ici par des raisons de confidentialité), ces deux corpus doivent représenter deux projets récents différents – mais pour lesquels nous supposons tout de même une « culture d'entreprise » commune – et ainsi nous permettre de réaliser des analyses semi-automatiques (elles-mêmes suivies de révision manuelles) pour mettre en évidence leurs points communs et leurs différences.

À des fins de comparaison, nous avons également constitué deux corpus *ad hoc*[7], de même langue (français) et de même taille, mais représentant des genres en théorie bien distincts : le premier est un extrait d'un manuel concernant les théories et les techniques des véhicules spatiaux, écrit au CNES par des experts pour des semi-experts, et le second est une collection d'articles extraits du quotidien *Le Monde*. Nous supposons que le manuel technique est plus proche par nature des spécifications, et ce à plusieurs égards (puisqu'il s'agit dans les deux

[5] Pour un parallèle entre sous-langages et langues contrôlées, voir en particulier Kittredge (2003).

[6] Ou en reconstruisant la grammaire du genre textuel.

[7] Nous utilisons ici le mot *corpus* dans un sens très lâche, puisque nous n'avons pas veillé à des questions aussi essentielles que la représentativité, dans la mesure où ces textes ne constituent pas notre objet d'étude, mais seulement des points de comparaison commodes.

cas de textes spécialisés à propos du domaine spatial, référant à des objets semblables). En revanche, le corpus de presse couvre de nombreux domaines et est nettement moins spécialisé.

4. Méthodologie

Notre objectif à ce stade est double : premièrement, montrer l'existence de spécificités linguistiques au sein de notre corpus d'exigences (par rapport aux deux autres corpus), qui confirmeraient l'hypothèse d'un sous-langage (ou genre textuel) et, deuxièmement, utiliser ces régularités pour proposer des règles de rédaction pertinentes et adéquates. Nous essayons donc, autant que possible, de combiner deux types d'analyses de nos données textuelles : une analyse quantitative, d'abord, destinée à chiffrer la fréquence d'apparition de certains phénomènes particuliers (ce qui nous permet d'en évaluer rapidement l'intérêt), et une analyse plus qualitative – requérant par conséquent une évaluation humaine –, ensuite, qui envisage certaines occurrences trouvées dans le corpus pour voir comment lesdits phénomènes se manifestent et comment ils devraient être envisagés par les règles de notre langue contrôlée (ce qui se traduit le plus souvent par une typologie distinguant les cas potentiellement litigieux de ceux qui ne semblent pas poser problème et par une identification des facteurs de risque).

S'agissant des phénomènes à considérer, nous souhaitons, là aussi, combiner les apports de deux approches bien établies en linguistique de corpus : l'approche *corpus-based* (qui, selon Biber (2009), présume la validité de formes et de structures linguistiques dérivées de la théorie linguistique) et l'approche *corpus-driven* (qui, toujours selon Biber (2009), est plus inductive, de telle façon que les constructions linguistiques elles-mêmes émergent de l'analyse d'un corpus), ainsi nommées par Tognini-Bonelli (2001).

Concrètement, dans une perspective *corpus-driven*, nous pouvons faire usage d'outils de traitement automatique des langues (en particulier de *text mining*) pour détecter automatiquement des structures récurrentes dans le corpus, et juger ensuite de leur pertinence pour notre recherche. Dans une perspective *corpus-based*, en revanche, il est nécessaire de fournir les hypothèses qui seront vérifiées par l'analyse du corpus ; dans notre cas, nous estimons que le meilleur point de départ possible se trouve dans les langues contrôlées déjà existantes[8], et en particulier dans celles qui sont dédiées à la rédaction technique. En effet, bien que nous puissions parfois leur reprocher de manquer de cohérence ou d'être approximatives, elles n'en sont pas moins l'œuvre de spécialistes du domaine, dont l'expérience est évidemment précieuse : en théorie, chaque règle se veut être la réponse à un problème précis, même si dans les faits, cela n'est pas toujours aussi simple et que la pertinence de certaines d'entre elles soit discutable. Nous nous attarderons dans cet article sur le *Guide for Writing Requirements,* publié par INCOSE (International Council on Systems Engineering) en 2011 et qui se présente comme étant l'état de l'art des règles de rédaction des exigences (indépendamment du domaine ou de l'industrie). Nous pouvons également distinguer entre les phénomènes directement visés par les règles des langues contrôlées (par exemple, l'une d'elles recommande de ne pas utiliser de pronoms – même si c'est sans doute le problème plus général de l'anaphore qui est ici envisagé) et ceux que nous avons extrapolés

[8] Ces langues contrôlées, à leur tour, s'inspirent entre autres de formules de lisibilité (DuBay, 2004) pour déterminer quels phénomènes sont les plus susceptibles de poser des problèmes de compréhension.

à partir de celles-ci (par exemple, puisqu'une règle interdit l'usage de la voix passive – qui permet d'éviter de mentionner l'agent –, nous avons supposé que les rédacteurs seraient davantage tentés d'utiliser le pronom *on*, qui permet d'utiliser la voix active sans être plus précis). Enfin, le choix des phénomènes que nous envisageons est aussi dicté par des raisons pratiques : tout d'abord, les langues contrôlées existantes étant pour la plupart destinées à la rédaction en anglais, nous avons dû sélectionner des phénomènes qui pouvaient être facilement transposés au français ; ensuite, puisque nous avons besoin d'outils de traitement automatique pour effectuer nos analyses – et parce que nous voudrions idéalement que les règles que nous proposons puissent être implémentées dans des outils d'aide à la rédaction –, nous avons dû nous assurer que de tels outils étaient effectivement disponibles (et si possible open-source) pour le français[9].

La figure 1 illustre schématiquement notre méthodologie, qui part des corpus et des règles existantes pour proposer de nouvelles règles, censément plus proches de la réalité.

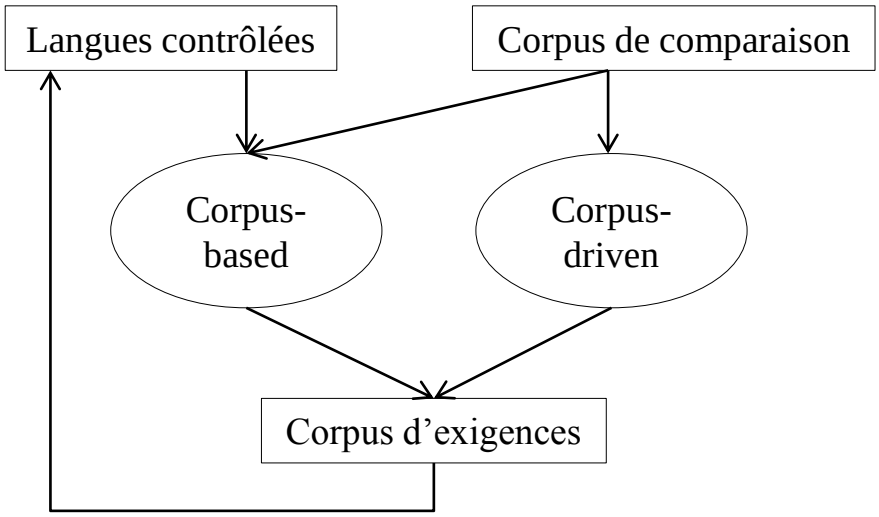

Figure 1 : Vers de nouvelles règles, inspirées des régularités en corpus

5. Résultats

Dans cette section, nous voudrions présenter brièvement quelques résultats obtenus lors de l'analyse en corpus de trois phénomènes illustrant les différentes facettes de notre méthodologie : un phénomène explicitement visé par les langues contrôlées (les pronoms), un phénomène extrapolé d'après les langues contrôlées (le pronom *on*) et un phénomène émergeant du corpus (le futur).

5.1. Les pronoms

Une règle du *Guide for Writing Requirements* décourage explicitement l'emploi des pronoms, précisant : "Repeat nouns in full instead of using pronouns to refer to nouns in other requirement statements. Pronouns are words such as 'it', 'this', 'that', 'he', 'she', 'they', 'them'.

[9] Ce qui exclut d'office l'analyse de phénomènes sémantiques trop complexes : une règle préconise par exemple qu'une exigence soit exprimée à un niveau de détails approprié par rapport à son niveau d'abstraction. Déjà difficile à appréhender pour un être humain, cette recommandation est évidemment hors de portée des machines actuelles.

When writing stories, they (sic.) are a useful device for avoiding the repetition of words; but when writing requirements, pronouns should be avoided, and the proper nouns repeated where necessary."

Un étiquetage morphosyntaxique effectué à l'aide de l'analyseur syntaxique Talismane (Urieli, 2013) montre que les pronoms (toutes catégories confondues) représentent environ 1,86 % des mots présents dans le corpus d'exigences, contre 2,93 % dans le manuel et 5,11 % dans les articles de presse. Il y a donc, dans les exigences, proportionnellement près de trois fois moins de pronoms que dans la presse (où la répétition est souvent vue, en français du moins, comme une faute de style) et une fois et demie moins que dans le manuel technique.

Si cette différence peut partiellement s'expliquer par d'autres raisons (les exigences étant plus courtes, la nécessité d'utiliser des pronoms se fait probablement moins souvent sentir), il n'en reste pas moins intéressant de constater que les pronoms semblent effectivement bien moins utilisés dans les exigences, ce qui pourrait constituer une régularité de genre et s'expliquer par la crainte de l'ambiguïté potentielle qu'ils représentent. Il est cependant également remarquable que les pronoms restent malgré tout présents : d'une part, certains sont inévitables (ou ne peuvent en tout cas pas être remplacés par un nom), tels que les pronoms relatifs ou les pronoms impersonnels ; d'autre part, certains ne semblent pas pouvoir prêter à confusion, tel le pronom personnel *il* dans l'exigence suivante : « Le paquet ne sera généré que s'*il* est activé par le LVC », où il ne peut se référer qu'à *paquet*. Si, comme conseillé par la règle envisagée plus haut, *il* devait être remplacé par le nom auquel il réfère, la phrase qui en résulte (« Le paquet ne sera généré que si le paquet est activé par le LVC ») paraîtrait très peu naturelle (sans que cela ne se justifie), voire pourrait prêter à confusion (s'agit-il bien du même paquet ?). De même, dans l'exigence « Le générateur de TC ne rejettera pas la création du PARAM_ID diagnostic si *celui-ci* est déjà défini à bord », l'utilisation d'un pronom démonstratif ne crée pas d'ambiguïté. Certains pronoms, en revanche, s'avèrent clairement problématiques : c'est le cas d'*Il* dans l'exigence « *Il* calculera aussi, à une fréquence paramétrable (ordre de grandeur 1 mois), la moyenne de mise en œuvre et la comparera à la moyenne maximum afin d'anticiper un problème éventuel », dont le référent ne peut pas être connu[10].

Il ressort de ces brefs exemples que remplacer systématiquement tous les pronoms par les noms correspondants n'est pas nécessaire, et serait même probablement contreproductif. Dans certains cas, néanmoins, l'usage d'un pronom devrait être fortement déconseillé (lorsque plusieurs référents sont possibles), voire interdit (lorsque le référent n'est pas présent). C'est probablement dans ce sens que la règle devrait être formulée (un pronom doit avoir un et un seul référent possible au sein de l'exigence, de préférence placé avant). En outre, comme déjà mentionné, on suspecte que c'est le problème de l'anaphore et de la cataphore (souvent sources d'ambiguïté) – plus que celui des pronoms – qui devrait faire l'objet d'une règle.

5.2. Le pronom on

Comme de nombreuses autres langues contrôlées (O'Brien, 2003), le *Guide for Writing Requirements* impose l'usage de la voix active[11] : "Use the active voice with the actor clearly

[10] Sinon en lisant l'exigence précédente, susceptible d'être déplacée ou supprimée par la suite. Chaque exigence devrait en théorie être autonome.

[11] Certaines langues contrôlées interdisent la voix passive, ce qui revient au même.

identified." Cette règle se justifie principalement par le fait que la voix passive permet beaucoup plus facilement l'omission de l'agent, ce qui se traduit généralement par un problème d'incomplétude des exigences. Les rédacteurs cherchent parfois volontairement à laisser l'agent relativement vague, pour plusieurs raisons (parce qu'il n'est pas encore connu précisément, pour que l'exigence paraisse moins directive, etc.) et nous pensons dès lors qu'ils peuvent être tentés de recourir à d'autres tournures pour y parvenir, avec notamment (en français) l'utilisation du pronom *on*.

Pour des extraits de seulement 53 000 mots chacun, nous en avons en effet relevé 158 occurrences dans le corpus d'exigences, soit étrangement beaucoup moins que dans le manuel (410, ce que nous expliquons par un désir de paraître aussi général que possible), mais plus que dans la presse (130). Ce nombre paraît à première vue relativement élevé pour un genre de texte dans lequel le flou – auquel ce pronom indéfini se prête particulièrement bien – est proscrit.

Lorsque l'on s'intéresse à ces occurrences, on constate qu'il peut prendre des valeurs très différentes. Certains usages semblent ne présenter aucun risque : « *On* se rapportera à DA14 pour la description de l'interface » (où *on* se rapporte de toute évidence au lecteur) ou « *On* ne listera ici que les TC associées aux familles F_EMETTEUR et F_MANOEUVRES » (où il réfère cette fois au rédacteur, tout en permettant un effacement énonciatif). Certains emplois permettent d'ailleurs de gagner en généralité : « *On* ne peut écrire et lire simultanément le même fichier » et « En conséquence *on* ne mixera jamais des clés C et M » (en prenant les valeurs de « personne » ou de « rien », il rend ces règles valables en toutes circonstances, pour tous les agents possibles). D'autres, en revanche, semblent superflus et trahissent peut-être une tendance à le surutiliser : « *On* considère que les autres TC sont associées à la famille F_AUTRES » (ici, *on* et la proposition principale pourraient tous deux être supprimés : « [Par convention,] les autres TC sont associées à la famille F_AUTRES ») ou encore « *On* prendra comme valeur d'origine du syndrome la valeur 0XFFFF » (qui pourrait être réécrit en « La valeur d'origine du syndrome sera 0XFFFF »). Enfin, certains exemples semblent réellement problématiques, dans la mesure où l'agent n'est pas connu : « *On* contrôlera les bornes MIN et MAX de la FT BDS […] associée à ce paramètre » et « *On* initialisera le modèle quand la batterie est en charge complète », ce qui est équivalent à une formulation à la voix passive dans laquelle l'agent ne serait pas précisé (on constate donc que le problème de l'incomplétude ne se limite pas à cette dernière). Nous pensons donc qu'une règle devrait par exemple soit interdire l'usage du pronom *on*, soit préciser dans quels cas il peut ou non être employé.

5.3. Le futur

Toujours pour des extraits de seulement 53 000 mots, une analyse des temps verbaux révèle que le futur de l'indicatif est beaucoup plus fréquent dans le corpus d'exigences (495 verbes, contre 2 671 au présent de l'indicatif) que dans le manuel (194 verbes au futur, 2 867 au présent) ou les articles de presse (156 verbes au futur, 2 652 au présent).

Ce constat peut s'expliquer par le fait que les exigences décrivent un système qui n'existe pas encore, mais *devra exister* plus tard tel qu'il est décrit au moment de la rédaction, d'où une tendance naturelle à employer le futur (« Le vidage des tables *sera* contrôlé par le CCC »). Là aussi, on peut penser que ce dernier participe notamment d'une stratégie visant à exprimer l'injonction inhérente aux exigences – une exigence devant nécessairement être suivie – tout en l'adoucissant ou en diluant la responsabilité (l'exigence ne devant plus être remplie immédiatement, mais à un moment futur qui n'est généralement pas précisé). Ce caractère

injonctif est d'ailleurs parfois rappelé par l'emploi de verbes modaux, en particulier *devoir* : « Le plan TC *devra* respecter les contraintes décrites dans DR20 ». Le problème réside donc ici dans l'absence d'indicateur temporel clair, ainsi que dans la possible confusion lorsque le futur et le présent sont employés alternativement, sans que l'on sache précisément pourquoi. Il serait donc préférable qu'une règle précise quels temps verbaux doivent être privilégiés et dans quels cas – le plus simple étant d'imposer le présent de façon systématique.

Ces deux derniers phénomènes (*on* et futur) et leurs manifestations dans le corpus d'exigences nous paraissent assez symptomatiques du conflit pouvant exister au sein des spécifications : d'un côté, la nature injonctive des exigences voudrait se manifester de façon parfaitement directe, ainsi que le préconisent les langues contrôlées visant la clarté ; de l'autre, les rédacteurs, étant des êtres humains s'adressant à d'autres êtres humains, cherchent malgré eux à l'atténuer par des formes linguistiques qui paraissent moins contraignantes.

6. Conclusion et perspectives

Dans cet article, nous avons voulu présenter la méthodologie que nous avons mise au point pour proposer une langue contrôlée destinée à la rédaction des exigences au CNES, dans le but d'éviter à l'avenir les risques liés à l'utilisation du langage naturel, qui peuvent avoir des conséquences importantes. En effet, bien que des langues contrôlées pour la rédaction technique existent déjà, nous jugeons que leurs règles ne sont pas applicables en l'état et demandent à être raffinées. Cette méthodologie repose avant tout sur l'hypothèse de l'existence d'un sous-langage (ou genre textuel) utilisé par la communauté des rédacteurs techniques lors de la spécification de systèmes spatiaux ; hypothèse que nous avons voulu vérifier par des analyses quantitatives. Elle s'appuie ensuite sur deux approches différentes, mais complémentaires : une approche dite *corpus-driven* (utilisant des outils statistiques) et une approche dite *corpus-based* (pour laquelle nos hypothèses de recherche nous sont principalement fournies par les règles présentes dans des langues contrôlées existantes), qui doivent nous permettre, grâce à des outils de traitement automatique du langage, de mettre au jour des régularités langagières qui serviront ensuite de base concrète à l'élaboration de nouvelles règles, plus proches cette fois des habitudes des ingénieurs. Nous avons enfin illustré cette méthode en analysant rapidement quelques phénomènes linguistiques simples qui se manifestent au sein du corpus d'exigences authentiques que nous avons constitué à cette fin.

Bien entendu, de nombreuses autres analyses restent possibles pour affiner les règles existantes ou en proposer de nouvelles, qui pourront constituer une langue contrôlée que nous espérons pouvoir à terme intégrer dans un outil de vérification semi-automatique, et ainsi faciliter la tâche des rédacteurs. Nous prévoyons également de constituer un corpus d'exigences provenant d'un domaine différent du spatial, ce qui nous permettra peut-être d'identifier des régularités relevant du genre des exigences en général, et non pas uniquement de celles écrites au CNES. Par ailleurs, nous voudrions vérifier nos hypothèses à propos de certains phénomènes linguistiques (voix passive, nominalisations) et leurs conséquences sur la clarté des exigences au moyen de questionnaires à destination des ingénieurs du CNES, dont nous pourrons également recueillir les avis.

Remerciements

Ce travail est effectué dans le cadre d'une thèse financée et activement soutenue par le Centre National d'Études Spatiales et le Conseil régional de Midi-Pyrénées.

Références

Bhatia V. K. (1993). *Analysing genre: Language use in professional settings*. London, Longman.

Biber D. (2009). Corpus-Based and Corpus-driven Analyses of Language Variation and Use. In *The Oxford Handbook of Linguistic Analysis*, 1st ed., Heine B. et Narrog H. éditeurs. Oxford University Press.

Carlson N. et Laplante P. (2013). The NASA automated requirements measurement tool: a reconstruction. *Innovations in Systems and Software Engineering*, vol. 10, no. 2, pages 77-91.

Clark P., Murray W. R., Harrison P. et Thompson J. (2010). Naturalness vs. Predictability: A Key Debate in Controlled Languages. In *CNL 2009 Workshop*, Marettimo (Italie), pages 65-81.

Condamines A. et Warnier M. (à paraître). Towards the Creation of a CNL Adapted to Requirements Writing by Combining Writing Recommendations and Spontaneous Regularities: Example in a Space Project. *Language Resources and Evaluation*.

DuBay W.H. (2004). *The principles of readability*. California, Impact Information.

Gaudin F. (1993). *Pour une socioterminologie: Des problèmes sémantiques aux pratiques institutionnelles*. Rouen: Publications de l'Université de Rouen.

IEEE (1990). Standard Glossary of Software Engineering Terminology. IEEE Standard 610.12-1990, 1-84.

International Council on Systems Engineering. (2011). Guide for Writing Requirements. Version 1.

Kamsties E. et Peach B. (2000). Taming ambiguity in natural language requirements. In *Proceedings of the Thirteenth International Conference on Software and Systems Engineering and Applications*, Paris (France).

Kittredge R. I. (2003). Sublanguages and Controlled Languages. In Mitkov R. editeur, *The Oxford Handbook of Computational Linguistics*, pages 430-447.

Kuhn T. (2014). A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, vol. 40, no. 1, pages 121-170.

Meyer B. (1985). On Formalism in Specifications. *IEEE Software*, vol. 2, no. 1, pages 6-26.

Nuseibeh B. et Easterbrook S. (2000). Requirements Engineering: A Roadmap. In *Proceedings of the Conference on The Future of Software Engineering*, New York (USA), pages 35-46.

O'Brien S. (2003). Controlling Controlled English. An Analysis of Several Controlled Language Rule Sets. In *Proceedings of EAMT-CLAW*, pages 105-114.

Pace G. J. et Rosner M. (2010). A Controlled Language for the Specification of Contracts. In *CNL 2009 Workshop*, Marettimo (Italie), pages 226-245.

Somers H. (1998). An Attempt to Use Weighted Cusums to Identify Sublanguages. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, Stroudsburg (USA), pages 131-139.

Temnikova I., Baumgartner Jr W. A., Hailu N. D., Nikolova I., McEnery T., Kilgarriff A., Angelova G. et Cohen K. B. (2014). Sublanguage Corpus Analysis Toolkit: A tool for assessing the representativeness and sublanguage characteristics of corpora. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1714-1718.

Tognini-Bonelli E. (2001). *Corpus Linguistics at Work*. John Benjamins Publishing.

Urieli A. (2013). Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit (PhD thesis). Université Toulouse II - Le Mirail.

Zhang Q. (1998). Fuzziness - vagueness - generality - ambiguity. *Journal of Pragmatics*, vol. 29, no. 1, pages 13–31.