A Methodology for Identifying Terms and Patterns Specific to Requirements as a Textual Genre Using Automated Tools

Maxime Warnier
CLLE-ERSS (UMR 5263)
Université Toulouse – Jean Jaurès & CNRS
Centre National d'Études Spatiales
maxime.warnier@univ-tlse2.fr

Anne Condamines
CLLE-ERSS (UMR 5263)
Université Toulouse – Jean Jaurès & CNRS
anne.condamines@univ-tlse2.fr

Abstract

As a step in a project whose final goal is to propose a Controlled Natural Language for requirements writing at CNES (Centre National d'Études Spatiales), we intend to build the grammar of the textual genre of the requirements. One of the main issues faced when analyzing our corpus is the (sometimes subtle) difference between the terms and syntactic structures pertaining to the genre and those linked to the domain (in our case, the development of space systems) – a difference that is generally not taken into account by automated tools. In this paper, we present a methodology aimed at detecting candidate terms and textual patterns specific to the genre by combining results obtained from a terminology extractor and a data mining tool with a validated resource in use for indexing documents at CNES. The results are then illustrated by a selection of examples from our corpus.

1 Introduction

This study is part of a wider project aiming at improving the writing of requirements[1] at CNES (Centre National d'Études Spatiales), the French Space Agency.

Indeed, the requirements (as well as the specifications, that is, the documents in which they are included) are mostly written in a natural language – in this case, in French –, and as a consequence they may sometimes contain well-known related problems, such as ambiguity and vagueness (Pace & Rosner, 2010). A Controlled Natural Language (CNL) is a possible solution to avoid or at least substantially limit these problems by setting constraints on the lexicon, the syntax or the semantics (Kuhn, 2014).

However, in order for this CNL to be actually applied, we believe that it should not be unnecessarily restrictive and, in particular, not too far removed from the way engineers are already used to write the documents – otherwise, they will probably merely ignore it. In other words, we wish to propose a CNL inspired by already existing data, following a corpus-driven and corpus-based methodology that we describe more in details in (Condamines & Warnier, 2014).

This methodology relies on the existence of a *textual genre*, which Bhatia (1993) defines as "a 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", as it is clearly the case for requirements writing (since it is a recurring task performed by employees working in similar companies), and in particular of a *sublanguage*, defined by Somers (1998) as "an identifiable genre or text-type in a given subject field, with a relatively or even absolutely closed set of syntactic structures and vocabulary". We were already able to provide some evidence in favor of this hypothesis (if not for all requirements, at least for requirements written in French at CNES) and we are now trying to build the grammar (that is to say the set of rules followed – consciously or not – by the speakers of this community to produce acceptable utterances) of this particular genre by semi-automatically analyzing specifications of two former projects.

In the present study, we will focus on the results obtained by a terminological extraction. More specifically, we will propose a method to sort them (as we are interested only in the terms pertaining to the genre, not in those pertaining to

[1] According to one of the definitions given by IEEE (1990), a *requirement* is: "a condition or capability that must be met or possessed by a system or system component to satisfy a contract, standard, specification, or other formally imposed documents".

the domain) and subsequently to use them as a filter to retrieve textual patterns belonging to the grammar of the genre. An example of similar work, based on collocations and n-grams, is given by the transdisciplinary scientific lexicon (Tutin, 2007).

2 Genre vs. domain

Although this grammar should ideally be independent of the field (aerospace industry, aeronautics, software engineering, etc.), in practice, the distinction is not so simple as regards specifications[2]. While some features are indeed inherent in the nature of the documents (because they describe something that does not exist yet, but will have to exist and to conform with the requirements, the use of the future tense and injunctions, for instance, are common), others, however, are closely related to the field to which belongs the future "object" being described. It may reasonably be assumed that the lexical level – since it directly refers to the object in question – is most significantly affected by the domain, but we cannot reject the hypothesis that syntactic structures too may differ from one field to another.

For that reason, if we want to define a terminology of requirements, we must keep in mind that the candidate terms proposed by the terminology extractors may actually belong either to the genre or to the domain. Unfortunately, although the possibility to filter terms by domain has already been highlighted as a user need (Blancafort et al., 2011), traditional extractors do not provide any means to distinguish *a priori* between genre and domain, because they are designed mostly for more didactic corpus, where the field matters much more than the genre (e.g. in order to establish the terminology in use in a company or in a knowledge domain). Furthermore, similar problems are to be expected when using other kinds of automated tools (such as data mining software), as they will also mix the two different types of words and terms.

Specifications are thus unusual, specialized corpora and they bring new challenges to terminology extraction in general. In particular, considering the fact that the candidate terms linked to the domain are probably more numerous than those linked to the genre, we want to find a way to exploit the results without a need for manually revising all of them. In the next section, we present the small experiment we conducted on our corpus of specifications as a possible way to reach this goal, but also to reuse these results to filter textual patterns identified by a text mining tool.

3 Methodology

3.1 Corpora

All the operations described hereafter were performed on two corpora of requirements in French extracted from several specifications provided by the CNES. (All tables and figures were removed from the requirements, because their automatic analysis would have been more difficult.) The first corpus concerns the project called "Pleiades" [3] (two very-high-resolution satellites for Earth observation) and is composed of nearly 120,000 words; the second corpus, related to the smaller project "Microscope"[4] (a microsatellite, whose main objective is to verify a physical principle), contains nearly 44,000 words. Although the requirements were written under similar circumstances and represent the same levels of specifications for the two projects, it is worth noting that Pleiades and Microscope have totally different scales and purposes. Consequently, the fields to which they relate are at least partially distinct.

3.2 Candidate terms

First of all, candidate terms for both corpora were extracted using the terminology extractor developed for the Talismane toolkit (Urieli, 2013); based on a syntactic analysis, it extracts only contiguous noun phrases. The first list we obtained (Pleiades) contained 1,551 candidates, while the second one (Microscope) contained 716 candidates (minimum frequency = 5).

Since they included candidate terms for the genre and for the domain (see section 2), and since we are interested only in the former, all the entries present in a list of terms used at CNES for indexing documents in their knowledge base were removed. This list of domain terms (used here as a "stop list") has been augmented for many years thanks to internal documents of various types and carefully validated by domain

[2] The distinction between *genre* and *domain* itself is actually far from trivial (Lee, 2001).

[3] https://pleiades.cnes.fr/en/PLEIADES/index.htm

[4] http://missions-scientifiques.cnes.fr/MICROSCOPE/

experts. We therefore assume that the terms that it contains are representative of the fields covered by the different projects conducted at CNES over the past years; furthermore, it is safe to think that it should not contain terms belonging to the genre of requirements, because they would not be helpful for indexation (since they are too general). After this step, only 1,355 entries remained for Pleiades (a difference of almost 200 entries) and 598 for Microscope (more than 100 candidates were thus discarded).

In order to remove even more candidate terms supposedly linked to the field, we decided to keep only entries present in both lists (Pleiades and Microscope). This resulted in a much shorter list of just 300 candidate terms (meaning 1,055 were exclusive to Pleiades and 298 to Microscope). This step makes sense because the specifications of Pleiades and Microscope are comparable at many levels, but also because, as already mentioned, the two projects are sufficiently distinct. Hence, whereas the first selection was useful to eliminate candidates related to the field at a more general level (e.g. "satellite" or "simulation"), here some of the candidates were not kept because they are more dependent to one of the two projects, and thus more specialized (e.g. "magnétomètre" 'magnetometer' or "masse interne" 'internal mass'). (However, because the corpus of specifications from Pleiades is almost three times larger than the other corpus, it is also probable that some terms, such as "priorité" 'priority', could have appeared in the Microscope corpus as well.)

Lastly, we proceeded to a manual revision of the remaining candidate terms to eliminate some entries that were obviously noise. The final list contains 267 candidate terms (to be compared with the original list, which would have contained over 1,850 different candidates, or almost 2,000 if the extraction had been performed on the two corpora as a whole). Interestingly, the terms seem to concern both functional requirements (e.g. "fonctionnalité" 'functionality') and non-functional requirements (e.g. "disponibilité" 'availability').

3.3 Textual patterns

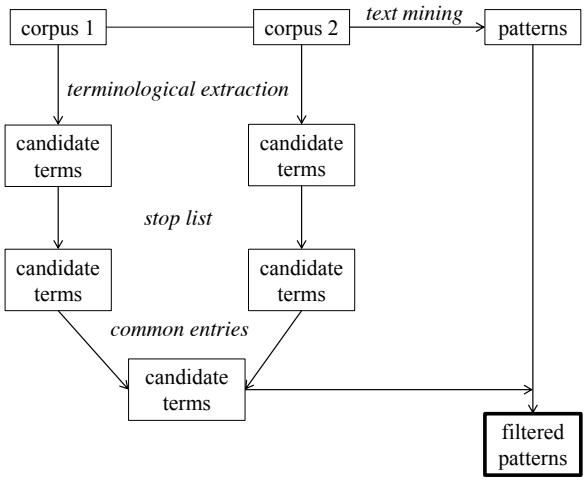
Of course, a grammar of genre should not be limited to the lexicon, as it would be the case with the results of the terminological extraction. We would like to identify recurring syntactic structures or, at least, frequent textual patterns[5] with the help of text mining tools.

For this purpose, we used SDMC (Quiniou et al., 2012) to retrieve patterns of *lemmas* (i.e. canonical forms of the words) frequent in the two corpora, such as "comme décrire dans le tableau" 'as describe in the table', appearing seventeen times in total. These patterns have variable lengths. Here again, the main problem is the huge number of results: almost 14,000 patterns were proposed, making a manual revision extremely time-consuming.

In order to reduce this number to a more reasonable proportion, we have decided to keep only patterns containing at least one of the remaining candidate terms (for the sake of simplicity, the noun phrases were reduced to their heads); indeed, we assume that the structures based on terms belonging to the genre are themselves more likely to be typical of this same genre. This restriction limited the number of patterns to approximately 6,000, among which "être connaître avec un [précision][6] meilleur que (number)" 'be know with a [precision] better than (number)', "être conforme au [format]" 'be consistent with the [format]" and "devoir respecter le [contrainte]" 'must respect the [constraint]'.

The list can be further reduced by focusing on patterns containing a verb. In this way, we consider an intermediary level between the lexicon and the discourse.

To conclude this section, the main steps of the process we described are represented by Figure 1.



[5] Patterns of this kind are the basis of the so-called "boilerplates" (Hull et al., 2005), which are basically fixed structures filled with variable elements at determined positions.

[6] The candidate terms are between square brackets.

Figure 1. Main steps of the proposed methodology.

4 Results

In this section, we briefly discuss some of the results we obtained after applying the process described previously.

4.1 Regarding terms

Some terms belonging to the space domain remain: initialisms ("ASH", "DGAPC"), terms too general to be useful for indexation ("mission", "centre de contrôle" 'control center'), terms of the field ("tuyère" 'nozzle', "calibration").

Others, by contrast, belong more to the genre. They may describe a need ("*besoin* de test+programmation+restitution" '*need* for a test+programmation+restitution') or the characteristics of the objet that is described ("*taille* du buffer temporaire+du paquet TM" '*size* of the temporary buffer+TM packet", "*durée* de désaturation+la manœuvre" '*duration* of desaturation+the manoeuvre'); they can specify expected functions ("*fonction* de gestion+filtrage" '*function* of management+filtering'); or they can be related to the management of the project: possible problems ("*défaillance*" '*failure*', "*défaut*" '*defect*'), necessary documentation ("*rapport* d'avancement+d'expertise" 'progress+expertise *report*'), validation ("*acceptation*" '*acceptance*', "*confirmation*", "*autorisation*" '*authorization*').

Some terms can belong either to the field or to the genre, depending on their modifier: "*date* de début du produit" 'starting *date* of the product' (genre) vs. "*dates* de début et de fin de vidage TM" 'starting and ending *dates* of the emptying of the TM' (field, because of the domain terms "vidage TM").

4.2 Regarding structures

The most frequent verbs in the patterns are: "être" 'to be', "devoir" 'must', "permettre" 'to allow', "mettre" 'to put', "prendre (en compte)" 'to take (into account)', "fournir" 'to provide', "pouvoir" 'to be able', "définir" 'to define', "passer (en mode+dans l'état)" 'to enter (a mode+a state)', "contenir" 'to contain', "donner" 'to give', "utiliser" 'to use', "gérer" 'to manage', "sélectionner" 'to select', "rejeter" 'to reject', "traiter" 'to process', "correspondre" 'to correspond', "générer" 'to generate', "décrire" 'to describe', "tenir" 'to hold', "exécuter" 'to execute', "vérifier" 'to verify', "calculer" 'to calculate'.

Some structures based on these verbs are typical of the corpus:

[Det N permettre de (V+deverbal noun)]: "le DUPC permettra de modifier localement les paramètres du calcul".

[Det N fournir Det N1 (à Det N2)]: "cette interface fournit les positions navigateur de l'instrument".

[Det N utiliser Det N2 (pour V)]: "le système GIDE utilisera le protocole FTP pour effectuer les transferts".

[Det N fournir (à Det N2) Det N3]: "le système de navigation fournira au système informatique central une réference de temps".

[Sur réception de cette TC, le LVC exécute la procédure de mise ON+OFF de Det N (, par l'envoi de commandes (sur+vers+à Det N3))]: "sur réception de cette TC, le LVC exécute la procédure de mise ON de la carte IOT sélectionnée, par l'envoi de commandes discrètes sur l'OBMU" (only in Pleiades).

[Det deverbal noun doit s'exécuter (conditions)]: "la consolidation du scenario de travail au CECT doit s'exécuter en moins de 15 secondes" (only in Microscope).

[Det N (avoir la capacité de+être (capable de+autorisé à)) traiter Det N2]: "le CCC doit avoir la capacité de récupérer et traiter 291 Mo de TM par jour".

These regular structures are therefore part of the grammar of the genre of requirements (at CNES).

5 Conclusion

As emphasized in section 2, specifications of space systems represent a particular type of corpus, because the terms of the domain and the terms of the genre are closely linked – making it difficult to automatically distinguish them. In section 3, we described the methodology we applied to keep only the terms belonging to the textual genre, using an existing resource (built for other needs) and a comparison between two corpora. This also allowed us to identify some structures (textual patterns) belonging to the grammar of the genre, which are used for writing functional requirements (describing expected functions) as well as for non-functional requirements (describing qualities or constraints applied to the system). The grammar could be refined thanks to existing guides to writing specifica-

tions that specify the various sections of the documents and the different types of requirements, which are likely to be expressed in different ways.

Nevertheless, it also appears that it is not always possible to draw a line clearly separating terms of the field and terms of the genre, since some terms may belong to both categories. In any case, the interpretation of the results remains dependent on the objective(s) being pursued.

Finally, we used this experiment as a proof-of-concept; before we can generalize it, we would have to ask for validation by experts (experienced writers). It would also be very interesting to compare our corpus to specifications written in another domain.

References

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

Blancafort, H., Heid, U., Gornostay, T., Méchoulam, C., Daille, B., & Sharoff, S. (2011). User-centred Views on Terminology Extraction Tools: Usage Scenarios and Integration into MT and CAT Tools. In *Conference "Translation Careers and Technologies: Convergence Points for the Future (TRALOGY)*. Paris, France: INIST.

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 B. Davis, K. Kaljurand, & T. Kuhn (Eds.), *Controlled Natural Language* (pp. 33–43). Springer International Publishing.

Hull, E., Jackson, K., & Dick, J. (2005). *Requirements engineering*. London: Springer.

IEEE Standard Glossary of Software Engineering Terminology. (1990). *IEEE Std 610.12-1990*, 1–84. http://doi.org/10.1109/IEEESTD.1990.101064

Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, *40*(1), 121–170.

Lee, D. Y. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the BNC jungle. Retrieved from http://ro.uow.edu.au/artspapers/598/

Pace, G. J., & Rosner, M. (2010). A Controlled Language for the Specification of Contracts. In N. Fuchs (Ed.), *CNL 2009 Workshop* (pp. 226–245). Marettimo: Springer.

Quiniou, S., Cellier, P., Charnois, T., & Legallois, D. (2012). What About Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics? In *International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'12)* (pp. 166–177). New Delhi, India.

Somers, H. (1998). An Attempt to Use Weighted Cusums to Identify Sublanguages. In D.M.W. Powers (Ed.), *NeMLaP3/CoNLL 98 : New Methods in Language Processing and Computational Natural Language Learning* (pp. 131–139). ACL.

Tutin, A. (2007). Modélisation linguistique et annotation des collocations: une application au lexique transdisciplinaire des écrits scientifiques. *Formaliser Les Langues Avec L'ordinateur: Actes Des Sixièmes, Sofia 2003, et Septièmes, Tours 2004, Journées Intex-Nooj*, *3*, 189.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Université de Toulouse 2 - Le Mirail, Toulouse.