Towards the creation of a CNL adapted to requirements writing by combining writing recommendations and spontaneous regularities : example in a Space Project

Anne Condamines, Maxime Warnier

Authors: Anne Condamines[1] (anne.condamines@univ-tlse2.fr), Maxime Warnier[1,2] (maxime.warnier@univ-tlse2.fr)

Title: Towards the Creation of a CNL Adapted to Requirements Writing by Combining Writing Recommendations and Spontaneous Regularities: Example in a Space Project

Affiliation: CLLE-ERSS (CNRS and Université Toulouse – Jean Jaurès) (Toulouse, France), CNES (France)

Abstract: The Quality Department of the French National Space Agency (CNES, Centre National d'Études Spatiales) wishes to design a writing guide based on the real and regular writing of requirements. As a first step in this project, the present article proposes a linguistic analysis of requirements written in French by CNES engineers. One of our goals is to determine to what extent they conform to several rules laid down in two existing Controlled Natural Languages (CNLs), namely the Simplified Technical English developed by the AeroSpace and Defense Industries Association of Europe and the *Guide for Writing Requirements* proposed by the International Council on Systems Engineering. Indeed, although CNES engineers are not obliged to follow any controlled language in their writing of requirements, we believe that language regularities are likely to emerge from this task, mainly due to the writers' experience. We are seeking to identify these regularities in order to use them as a basis for a new CNL for the writing of requirements. The issue is approached using natural language processing tools to identify sentences that do not comply with the rules or contain specific linguistic phenomena. We further review these sentences to understand why the recommendations cannot (or should not) always be applied when specifying large-scale projects.

Keywords: controlled natural language • corpus linguistics • requirements • technical writing • textual genre

1 Introduction

The study presented in this article was conducted with a view to improving the writing of requirements at CNES (Centre National d'Études Spatiales). Our ultimate goal is to propose a Controlled Natural Language (CNL) well adapted to this aim by taking into account both the recommendations proposed by existing writing guides and CNLs, and the linguistic regularities appearing in the specifications used at CNES.

The CNES and our laboratory have been collaborating for several years on questions concerning terminology, text management and the study of risks related to the use of language (Condamines 2010). As linguists, we propose methods and results based on a corpus linguistics approach, assisted by tools such as parsers, statistical analyzers, terminology extractors, concordancers or scripting languages. More recently, we were approached on the specific problem of writing requirements.

The CNES is the French space agency and, as such, is responsible for designing space systems. Therefore, it has to draft specifications (that must clearly and precisely describe its needs) which are intended for companies that respond to the bids; in turn, it also responds to bids from other scientific, commercial or military partners. The Quality Assurance Sub-Directorate of CNES, however, is aware that these specifications are not always clear, and that as a result there may be divergent interpretations, leading to additional costs, delays or even litigation (indeed, since requirements are part of the contract clauses, we may consider that they belong to what Kurzon (1997) and other authors call the "language of the law").

Two kinds of documents were used in the present study. On the one hand, two existing CNLs (the *Simplified Technical English* proposed by the AeroSpace and Defence Industries Association of Europe and the *Guide for Writing Requirements* issued by the International Council on Systems Engineering), and on the other a subset of the specifications of the space project Pleiades. Each of these two document types corresponds to one of the two approaches we aim to combine: prescription and spontaneous regularities.

The point of view underlying our approach is twofold. First, we consider that guides for technical writing are not fully adapted to the real writing process: they are sometimes too constraining, and sometimes insufficiently so. In most cases (and especially when they are designed for firms or organizations), they are not written by linguists but by domain experts who adopt a prescriptive stance based in part on traditional readability formulas and on their own experience as specialized writers. This is the case for example in the field of air traffic control where the ICAO (International Civil Aviation Organization) phraseology is written by controllers (Lopez et al. 2013). However, even though these guides are not always adapted to the reality of language use, they can constitute a good starting point – precisely because they are based on the experience of the domain experts. Secondly, we consider that requirements writing belongs to a *textual genre* that can be described in terms of situational and linguistic regularities. We think that these two approaches (prescriptive and descriptive) are not opposed but can be used together in order to design a CNL adapted to requirements writing.

The paper describes the first step of the project and focuses mainly on the feasibility of the method. It is organized in three parts. In Section 2, we describe in detail our project to combine prescription and description of spontaneous regularities. We situate the notion of spontaneous regularities close to that of *textual genre* and explain that the global perspective combines two concepts from the French Linguistic School of Rouen, *normalisation* and *normaison*. In Section 3, we present the methods and tools used to analyse the requirements corpus, in order to propose a linguistic diagnosis by comparing the requirements with the recommendations of the CNLs and by spotting regularities. In Section 4, we present some preliminary results of the study.

2 Combining normaison and normalisation

The two terms *normalisation* ("normalization") and *normaison* ("norming") were proposed by Guespin in the context of terminology. They were then used by researchers in Socio-terminology; Gaudin, among them, precises that:

"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."[1] (Gaudin 1993:173)

[1] "Analysis would gain by contrasting two normative processes: norming, which is an intrinsic feature of spontaneous language activity, and normalization, which is a conscious and planned activity."

The morphological similarity of the two terms shows that in both cases, it is the norm that is at issue. In the first case, however, the norm is considered from a descriptive and non-conscious point of view, whereas in the second, it is considered from a prescriptive and conscious point of view. In both kinds of norms, speakers need to follow the rules in order to be identified as belonging to a speech community (that is, a community of people who share rules about 'when' and 'how' to speak (Hymes 1967:54)).

2.1. Normalisation within terminology and discourse

Within terminology, normalisation is accomplished by authorized bodies composed of domain experts (assisted by terminologists) who decide which terms have to be used in order to designate concepts with a predetermined meaning. The idea underlying this process corresponds to the one proposed by Eugen Wüster, a Viennese engineer who created in the 1930s the "General Theory of Terminology" (Wüster 1968). In the hope of controlling language difficulties (ambiguity, vagueness, etc.), especially in multilingual situations, he proposed to build lists of terms and definitions by domain. This implies that terms are considered as different from general language words, because they are assumed to be monosemic in all linguistic or situational contexts; in other words, their meaning is fixed once and is assumed not to vary. This point of view has been called into question particularly since the 1990s, mainly for the following reasons (Condamines 1995):

- As it is anchored only in communicative needs, and highly different from a linguistic approach, it isolates the field of terminology, which is a theoretically untenable position.
- It quickly became clear that the terms and even the meanings proposed by terminological standards do not correspond, in most cases, to those used in technical documents.
- With the development of Natural Language Processing, it was easy to compare standards with terms used in specialized corpora. Moreover, a large number of tools dedicated to terminology (extraction of terms or conceptual relations) have been designed.
- Electronic document management is now widely developed and requires resources corresponding to real uses.

As a result, many researchers have criticized the Wüsterian viewpoint and have proposed new approaches, based on how terms are really used in specialized discourses. Examples are the Communicative Theory of Terminology (Cabré 1999), Textual Terminology (Pearson 1998), Socio-cognitive Terminology (Temmerman 2000) and Socio-terminology (Gaudin 1993). All these approaches consider that terms may vary, partly because they are words, but also because they have their own specificity due to the fact that they occur in specialized contexts (Condamines 2010; Freixa 2006).

CNLs, even if generally less normative than terminological norms, were partly defined with the same aim as the General Theory of Terminology, that of limiting language difficulties, at least when they were designed for firms. Hence, most CNLs designed for technical writing contain glossaries with forbidden or authorized terms, and acronyms (they are very frequent in these kinds of texts and need to be defined somewhere to be understood). The risk with CNLs is the same as with terminological norms, in that they can be too remote from language use and very difficult to apply (in other words, too many standards may kill standards).

In line with the evolution sketched above in the field of terminology, we propose to build a CNL not only with a view to "normalizing", but also by taking into account spontaneous regularities, that is, based on "normaison".

2.2. Normalisation, CNLs and readability formulas

The study of different CNLs highlights some recurring recommendations, inspired by readability formulas. As described by DuBay (2004), readability studies started in the late 19th century. They were first focused on general language, educational perspectives and based on psychological approaches (see for example (Flesch 1948)) and later more on communication studies. As pointed out by Klare (1976), the tests used for evaluating the formulas were based either on the selection of one meaning among others by readers after their reading task, either on the use of cloze tests ("a cloze test uses a text with regular deleted word […] and requires the subject to fill in the blank") (DuBay 2004:27). All the results were treated by statistical methods in order to validate or invalidate the supposed characteristic of a readable text. There have been only a few studies about the

efficiency of these simplified languages, such as Simplified English (Shubert et al. 1995; Chervak et al. 1996), and the results about Simplified English are not always significant (Stewart 1998).

If some frequent proposed formulas seem intuitively relevant (such as sentence or word length), others are more questionable, particularly when they concern not only the form but also the content of the texts. The first studies on readability have been criticized for their violation of knowledge about "reading and the reading process" (Bruce et al. 1981) or for their statistical bases; plenty of studies have addressed this issue (Klare 1976; Bruce et al. 1981; Gilliland 1972).

Nevertheless, readability formulas have been naturally integrated in the definition of CNLs. When defined for specific domains, the CNLs have been built by combining the current knowledge about readability and the knowledge of experienced writers. Therefore, CNLs recommendations may be considered as norms inspired by the historical characterization of readability adapted by domain experts.

As we will see later, our proposition is to build a new method for building a more adapted CNL by taking into account the characteristics of existing CNLs and the "spontaneous" writing of the domain experts.

2.3. Normaison

In the socioterminology school, *normaison* is defined as follows:

"Tout ensemble langagier permettant l'intercompréhension comporte ses normes systémiques : c'est cette logique que nous appelons « normaison »."[2] (Guespin 1993:217)

This definition expresses the fact that when speakers communicate regularly, they unconsciously regulate the way they speak and generate spontaneous regularities. It is exactly the same idea that underlies the notion of *textual genre*, regardless of the author's position and origins:

"Each separate utterance is individual, of course, but each sphere in which language is used develops its own *relatively stable types* of these utterances. These we may call *speech genres*." (Bakhtin 2004:60)

"A textual genre is 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." (Bhatia 1993:13, following Swales's work)

Note that the notion of textual genre is not always properly distinguished from that of *sublanguage*. See for example the definition by Somers:

"A sublanguage is an identifiable genre or text-type in a given subject field, with a relatively or even absolutely closed set of syntactic structures and vocabulary." (Somers 1998:131)

Other authors such as Kurzon (1997), Temnikova (2012) and Kuhn (2014) have noted this point. Historically, the most important difference is that the notion of sublanguage was proposed by Harris from a mathematical and distributional perspective (Kittredge and Lehrberger 1982), while that of textual genre comes from a more literary and historical approach (Bakhtin 2004), a sociolinguistic approach (Firth 1957; Bhatia 1993; Swales 2004) or even a corpus linguistic one (Biber 1988). Overall, one of the most important characteristics is that linguistic regularities are associated with speakers from the same speech community; here, this community is composed of several engineers working together on the same project and responsible for writing the necessary specifications (which naturally leads them to communicate frequently).

Two features characterize genres, whatever the definition:

- A recurrent communicative situation generates speech regularities.
- Speech regularities may be used in order to recognize which communicative situation is concerned.

This second point takes into account both lexical (terms) and morpho-syntactic specificities in order to define a grammar of the genre (in our case, that of requirements).

We can consider that normalisation corresponds to a top-down approach to norms, while normaison corresponds to a bottom-up one. Our aim is to combine these two approaches in order to design a CNL that will be well adapted to requirements writing. This involves spotting the regularities in existing requirements (to build the grammar of this genre), guided by the recommendations appearing in CNLs. In fact, we consider that the recommendations are relevant in their aims but not realistic in their implementation.

[2] "Any body of language that enables mutual understanding possesses its own systemic norms: it is this logic that we will call 'norming'."

2.4. Analysis of existing requirements guided by existing norms

This part presents in greater detail our method for combining normalisation and normaison. To achieve this, we use the two well-known corpus linguistics approaches described by Tognini-Bonelli (2001), the corpus-driven and the corpus-based approaches:

"The term corpus-based is used to refer to a methodology that avails itself of the corpus mainly to expound, test, or exemplify theories and descriptions that were formulated before large corpora became available to inform language study." (Tognini-Bonelli 2001:65) In the corpus-driven approach, however, "the theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus." (Tognini-Bonelli 2001:85)

So, the corpus-based approach corresponds to a top-down point of view, since assumptions are first made and then tested in the corpus, while in a corpus-driven (bottom-up) approach, the specificities of the corpus are first taken into account. In fact, the two approaches are very often complementary since, in a corpus analysis, there is a constant combination between the specificities of the corpus, the linguistic hypotheses interpreting these specificities, and verification of these hypotheses.

By adopting a corpus-driven perspective, the present study aims to contribute to the grammars of genre. According to grammars of genre, a textual genre can be characterized by linguistic regularities; data mining methods can be used to highlight these regularities. The tools developed generally combine statistical techniques with morpho-syntactic or sometimes semantic features (Poudat 2003; Biber 2006; Nishina 2007; Biber 2009). In the case of SubCAT (Temnikova et al. 2014), the texts belonging to the genre being studied are compared with a representative corpus of the same language. (Unfortunately, such a corpus does not exist for French for the moment.) Within a corpus-based perspective (McEnery and Wilson 1996), we have used the CNL recommendations – either directly or in order to understand the purpose targeted by each of them. Based on this analysis, the aim is to identify how (by which linguistic choices) this purpose is (or may be) taken into account from a discursive point of view.

To take an example (developed in more detail below), it is clear that injunction underlies requirements. The main objective in requirements is to state what the CNES wants the contractor to do. From a discursive perspective, however, linguistic patterns that express a command in a direct manner may (e.g. with the imperative) – at least in French – sometimes be felt as unpleasant. So in order to mitigate bald-on-record commands, speakers use different strategies such as the future tense.

In the following sections, we show that our proposal concerning the feasibility of building a CNL based both on normalisation and normaison is applicable. With the corpus-based perspective, several examples of the way direct recommendations (or the purpose underlying them) appear in the real requirements are given. This initial analysis should enable a double diagnosis. First we aim to show how large the gap between the writing norm and real use may be. Second, we also want to show that the "spontaneous" writing of requirements generates its own regularities.

3 Analysis of requirements: methods and tools

As we have seen, at this stage of the study, the aim is to see if our hypothesis concerning the feasibility of combining existing CNL and corpus regularities is valid. In this section, we describe the corpus of requirements we built and the CNL we used as references. Then we present the linguistic phenomena selected for study in these CNLs.

3.1. Description of the corpus

A subset of the specifications of an Earth observation satellite called Pleiades (launched in 2011) was obtained from the CNES to represent specifications of a large project typical of this space agency. These specifications were chosen because they were quite easily available and because they represent most of the levels of the classical product tree: system, segment, equipment and interfaces. They include both functional and non-functional requirements (which are not properly distinguished within the documents). Typical writers at CNES (including those involved in this project) were hired after they graduated from a French engineering school (5-

year program), where requirements writing is only little taught, and then formed inside the company. The number of authors per document ranges from 2 to 10, and all of them were validated by the project manager.

From these specifications, we extracted the requirements, that is to say only those parts that play the role of contractual obligations between the CNES and its subcontractors. Requirements should not contain unnecessary information, such as examples or comments, and they are intended to be autonomous, i.e. they are supposed to be understandable even without knowledge of the textual segments that precede or follow them. In the specifications we were given, the requirements were easily identifiable because they were framed by specific tags.

The requirements are all written in natural language, but some also contained tables or diagrams (which were removed, since they cannot be analyzed automatically). In theory, they should be fully understandable even without those figures (in particular thanks to captions) – but in practice, this is not always the case.

The resulting corpus is composed of 1,142 requirements (nearly 53,000 words – understood here simply as sequences of characters between spaces or other punctuations marks) in French. Although it is a small corpus, it should allow us to find utterances of the phenomena (linguistic forms and structures) that we intend to analyze.The final corpus that we will use for our future analyses will be composed of 3,595 requirements (163,000 words) from two different projects and should therefore allow comparisons between projects, to ensure a greater representativeness.

3.2. CNLs and rules

Many guides for writing requirements exist, and most of them were designed to avoid undesirable properties of unrestricted natural language (which "brings with it a host of well-known problems" (Pace and Rosner 2010)), such as ambiguity, polysemy, vagueness, and so on (Zhang 1998; Condamines 2010). Several other solutions have been proposed to improve the quality of requirements: natural language processing tools for (semi-)automatic verification (Carlson and Laplante 2013; Barcellini et al. 2012; among many others), "boilerplates" (fixed structures filled with variable elements at determined positions) or even more formal, univocal languages (Meyer 1985). The first solution is an aid offered to the users during or after the writing (downstream work); the second is simple to use but has limited expressiveness; the third one requires the users to actively learn a new language before they can start writing the specifications (upstream work) and, as a consequence, they are more demanding (especially if the language is far from the natural language). Of course, they can be combined.

At this step of the study, the studied CNLs correspond to human-oriented CNLs, i.e. read and interpreted by humans beings. But, since our aim is to propose new recommendations, we also want to anticipate the way of controlling their applicability and their automatic verifiability. Hence, our project is located upstream of the writing process but, if successful, we hope it will help improve the requirement engineering methods.

We selected a few rules from two different CNLs: (1) the well-known ASD *Simplified Technical English* (AeroSpace and Defence Industries Association of Europe 2007) (from now on, ASD STE), "an international specification for the preparation of maintenance documentation in a controlled language", and (2) the *Guide for Writing Requirements* proposed by the International Council on Systems Engineering (2011) (from now on, INCOSE). The aim of the latter guide is presented as follows: "to draw together advice from a variety of existing standards into a single, comprehensive set of rules and objectives". As a result, some rules from ASD STE and INCOSE may be (at least partly) similar. INCOSE is also quite general since it "is intended to cover the expression of requirements from across disciplines". It is therefore intended for engineers who write or review requirements.

Both can be clearly considered human-oriented or "naturalist" controlled languages (as opposed to the "formalist" approach) (Clark et al. 2010), whose goal is to facilitate human-to-human communication (Wyner et al. 2010). Like many other controlled natural languages aimed at improving communication among humans, their main purpose is to ensure that the message written in natural language has only one possible interpretation. It is worth noting that this conception of natural language is very different from that adopted by linguistics.[3] It

[3] According to Jakobson (1960), for example, the referential function, which is the closest to the one consisting in transmitting information, is only one among the six functions of language.

can be reasonably assumed, however, that by establishing guidelines in narrowly-defined situations, it may be possible to limit (if not to remove completely) the inherent difficulties linked to natural language.

INCOSE has the four characteristics of controlled natural languages proposed by Kuhn (2014), since it has one base language (English), it is a constructed language, it sets constraints on the vocabulary, the syntax and the semantics, and the resulting textual requirements are still understandable by English speakers. It is not a mere style guide, because the recommendations are real rules, not hints – even if the authors admit that "rules have to constantly be adapted to particular situations". All of them are followed by objectives that explain why the rules are useful. Among the main "objectives for writing requirement statements" are singularity, completeness, necessity, comprehensibility, concision, precision and non-ambiguity. These recommendations are translated into instructions which can be either direct or rather vague (e.g. "Express the level of detail appropriate to the layer in which the requirement lives"). We selected several of these instructions and analyzed our corpus to see how often they appear.

Because the phenomena we chose to observe are quite general (i.e. not highly language-dependent), we assume that most of the conclusions we propose for French are valid for English as well. In fact, INCOSE, while written in English and mainly based on older English guides, sometimes gives examples in French.

Since it was not possible to check the conformity of the requirements to all the recommendations proposed by ASD STE or INCOSE (partly because the study is still in its initial stage, and partly because several of the recommendations cannot be verified in an automated manner (e.g. "When a requirement is related to complex behaviour, refer to the supporting design or model" or "Group related requirements together"), we decided to focus on a selection. The manifold phenomena we intend to examine belong to three categories, all combining corpus-based and corpus-driven approach: (1) explicit rules in CNLs, (2) phenomena extrapolated from the CNLs and emerging from the corpus, and (3) phenomena emerging from the corpus and explained by the genre characteristics.

3.2.1. Explicit rules in CNLs

In this subsection, we examine some rules stated in INCOSE (combinators, pronouns) or ASD STE (length of sentences), before comparing them to our corpus to check whether they can be applied literally; the aim is to see if they are already suitable for the writing of requirements at CNES, or if they need to be adapted – or even rejected.

The first rule from INCOSE that we chose to compare with the corpus is called "Singularity/Propositionals" and states that "combinators" must be avoided: "Combinators are words that join clauses together, such as 'and', 'or', 'then', 'unless'. Their presence in a requirement usually indicates that multiple requirements should be written." Nevertheless, some of them are still present in the examples of "acceptable" specifications; this paradox suggests that these "combinators" cannot always be avoided.

The second rule is called "Completeness/Pronouns" and states that it is better to repeat nouns in full, rather than using pronouns to refer to nouns in other statements: "Pronouns are words such as 'it', 'this', 'that', 'he', 'she', 'they', 'them'. When writing stories, they 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." However, there is no indication about the conditions required for this repetition to be "necessary"; we can merely infer that the aim is to avoid problems due to anaphora resolution. Besides, in the only example given by INCOSE[4], the ambiguity lies in a determiner, not in a pronoun.

Although this rule is not present in INCOSE (which simply recommends "concise" requirements), ASD STE, like many other guides for technical writing, imposes a word limit for each sentence (probably because it is believed that longer sentences are harder to process, especially for non-native readers). This limit (usually around 20 words for English) depends here on the type of text: 20 words for procedural texts, or 25 words for descriptive texts (since, unlike INCOSE, ASD STE is not intended for writing requirements, but maintenance documentation). Besides the fact that these numbers of words seem quite arbitrary, counting words in a sentence

[4] "The controller shall send the driver's itinary (sic) for the day to the driver" must be preferred to "The controller shall send the driver his itinary (sic) for the day".

is far from trivial. ASD STE therefore provides many rules specifying how to do so[5]. In effect, the main advantage of such rules is that writers are more likely to split their instructions into several sentences in order to respect the rules.

In brief, we can already point out that these three rules are very general and seem way too restrictive, and that their justifications (if present) are evasive.

3.2.2. Phenomena extrapolated from the CNLs and emerging from the corpus

One of the most common rules in naturalist CNLs concerns the prohibition of the passive voice (O'Brien 2003), either by forbidding it or by imposing the active voice (as is the case in both ASD STE and INCOSE). There are mainly two reasons for this rejection: first, the active voice is preferred because it is considered more "canonical" (and easier to understand) and, secondly, since the passive voice allows omission of the agent[6], it can be a way to escape the responsibility of the instruction (since no one is clearly identified as the agent).

Nonetheless, we hypothesize that technical writers are naturally tempted to find other formulations to avoid the agent, even when prompted or forced to use the active voice. We assume that they tend to use the French pronoun *on* (whose closest English equivalents are the indefinite pronoun *one* and the generic *you*). Although formally a third-person singular subject pronoun, *on* can be used instead of any other subject pronoun (Bouquet 2007; Malrieu 2007) and, thus, is a convenient way to avoid specifying the human agent (for example, because the agent is not yet known). No rules in the two CNLs we studied (nor, to our knowledge, in any existing CNL) address this linguistic form and the way it can substitute for the passive voice.

3.2.3. Phenomena emerging from the corpus and explained by the genre characteristics

Lastly, it is interesting that many requirements in the corpus are expressed with a future tense, not with the present tense (even though the latter is more general). We can reasonably assume that the main reason why the future is used is because the requirements are not to be met immediately when written, but at a later, unidentified time: one very particular aspect of requirements is that they describe an object (in our case, a satellite) that does not yet exist, but will have to exist in the future in the exact same way as the engineers have designed it. Therefore, an essential condition of the future is that it can be used to express the deontic modality – which is the case, as noted by Stage (2002).

ASD STE lists the future tense as one of the authorized tenses in procedures, but imposes the imperative form for every instruction, adding that "less direct forms of instructions leave confusion as to whether something: must be done, or is already done, or must be done in the future by someone else". It is true that the imperative is the canonical grammatical mood for injunctions, but the future tense, like for instance some auxiliaries or adverbs, can be as strict – and sometimes even stricter – than the imperative. In fact, the strength of the order (that is, the exercitive speech act (Austin 1975)) depends first of all on the relation between interlocutors (Stage 2002; Le Querler 2004); in the present case, the engineers responsible for the implementation of the system have a professional obligation to follow it. Moreover, as pointed out by ASD STE, the agent in imperative instructions is clearly identified (i.e. the reader), while it is not necessarily expressed in sentences with the future tense. For this reason, we consider that it is relevant to analyse these sentences in conjunction with the pronoun *on*, briefly described above.

The methodology we have applied (illustrated by Figure 1) allows us to combine the corpus-driven and the corpus-based approaches, since we take into account phenomena that were identified as typical of the corpus (*on*, future verbs) and/or were addressed by some CNLs (conjunctions, pronouns, long sentences). This evidences the fact that these approaches can be complementary for the analysis of the corpus.

[5] But, once again, these rules can be difficult to understand or apply; one of them, for instance, states that "when you count words for sentence length, each word in a hyphenated group counts as a separate word unless it is a prefix", which implies a morphological analysis of the words to be properly applied.

[6] Compare: "The agent does the operation" (active voice) *vs.* "The operation is done [by the agent]" (passive voice). In the first case, the sentence would not be grammatical without its subject; in the second case, the agent is optional.

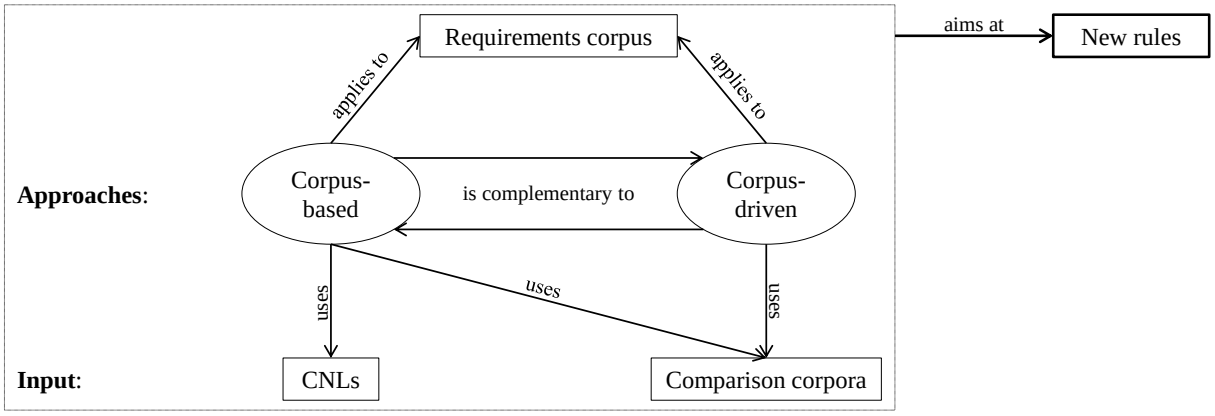

Fig. 1 Methodology: from phenomena to rules

3.3. Tools and resources

Several tools were used to perform the tasks described in Section 4. Syntactic analysis was performed using Talismane (Urieli 2013), an open-source parser developed in our laboratory, while the open-source corpus processor Unitex (Paumier 2011) was used for sentence chunking, and the concordancer TXM to find occurrences. Short handmade Perl scripts were written for other needs (extraction of the requirements, detection of long sentences, and so on).

We also compared our corpus to two other corpora (reduced to the exact same size): (1) a handbook written by experts from the CNES for a course about techniques and technologies used for building and operating spacecraft (e.g. a chapter concerns optical telecommunications and include considerations about their importance), intended for semi-experts, and (2) some articles from the French national newspaper *Le Monde*. This comparison aims to situate our corpus in relation to a technical corpus (in the same domain, written in the same firm) and to a "generic corpus", i.e. one not linked to a specific domain. As previously mentioned, there is unfortunately no representative corpus for French.

4 Preliminary results of the corpus analysis

In Subsection 4.1, we present the results concerning the frequency of conjunctions, pronouns, long sentences, *on* and future verbs in our corpus. In Subsection 4.2, we propose a selection of examples that break the two rules from INCOSE and try to classify them according to their necessity (mandatory, useful or undesirable) and also review examples of requirements containing the pronoun *on* or future verbs.

4.1. Quantitative analysis

Thanks to the syntactic analysis (useful for disambiguating the forms that may belong to different parts of speech), we were able to retrieve all the occurrences of the so-called combinators (since no exhaustive list was given, we looked for all coordinating and subordinating conjunctions) and all the pronouns in the corpus. As can be seen from Table 1, both are numerous, suggesting that they are common in unrestricted natural language.

Table 1 Number of conjunctions and pronouns in the three corpora[7]

Corpus	**Conjunctions**			**Pronouns**
	Coordinators	**Subordinators**	**Total**	
Requirements	882 (1.66%)	365 (0.69%)	1247 (2.35%)	986 (1.86%)
Handbook	1455 (2.75%)	442 (0.83%)	1897 (3.58%)	1554 (2.93%)
Newspaper	1274 (2.40%)	579 (1.09%)	1853 (3.50%)	2710 (5.11%)

[7] Percentages indicate the number of occurrences in relation to the total number of words.

Still, they are much less frequent in requirements than in the other two corpora, handbooks and newspapers. This is particularly clear in the case of pronouns, which are nearly three times more frequent in newspapers (where repetition is seen as an error of style in French) than in requirements (which are usually much shorter). Similarly, coordinating conjunctions are more frequent in the handbook, which has a didactic purpose (meaning that the links between propositions must be more clearly expressed), and subordinating conjunctions are used more commonly in the newspaper, because complex sentences are representative of a more elaborated style. We believe that such a marked difference is an argument in favor of our initial hypothesis that regularities spontaneously arise in daily practice, and that requirement writing can be considered a textual genre, even when not taught as such.

We also considered the length of the sentences composing the requirements. The results of our measures are shown in Table 2.

Table 2 Length of sentences in the three corpora

Corpus	**Sentences**	**Sentences with more than 25 words**	**Average number of words per sentence**
Requirements	4859	350 (7.2%)	11
Handbook	3456	591 (17.1%)	15
Newspaper	2201	839 (38.1%)	24

Once again, significant differences exist between the three types of documents: sentences tend to be shorter in requirements, and much longer in newspaper articles. However, long sentences are not uncommon in the requirements corpus; there is even one unusually long sentence containing over 70 words.[8]

Table 3 shows the frequency of the pronoun *on* in the three types of texts. A distinction is made between sentences where *on* is followed by a verb in the present tense and sentences where it is followed by a verb in the future.

Table 3 Number of *on* in the three corpora

Corpus	**On**	***On* + present verb**	***On* + future verb**
Requirements	158 (100%)	78 (49.37%)	78 (49.37%)
Handbook	410 (100%)	359 (87.56%)	42 (10.24%)
Newspaper	130 (100%)	83 (63.85%)	9 (6.9%)

Surprisingly, *on* is extremely frequent in the handbook, where examples tend to show a desire to be rather general ("pour augmenter le débit, on a intérêt à augmenter la puissance du laser" ["to increase the flow, one should increase the power of the laser"]) as well as the use of set phrases ("On remarque que…" ["It can be seen that…"]). A syntactic construction that is highly specific to the requirements, in contrast, is "*on* + future verb", which is much less frequent in the handbook and even less so in the newspaper.

Lastly, Table 4 gives an overview of the number of present verbs, future verbs, imperative verbs and conditional verbs.

Table 4 Number of present, future, imperative and conditional verbs in the three corpora

Corpus	**Present**	**Future**	**Imperative**	**Conditional**
Requirements	2671	495	0	11
Handbook	2867	194	0	21
Newspaper	2652	156	0	100

[8] "Si la différence (en valeur absolue) entre les dates de fin de lecture de deux fichiers, lus sur tranche de COME M - canal TMI i et sur tranche de COME N - canal TMI j, est inférieure à OPS_DELAI_INTER_FIN_LEC secondes, alors il est interdit d'enchaîner (lecture enchaînée) par la lecture de la tranche de COME N sur le canal i et de la tranche de COME M sur le canal j." ["If the difference (in absolute terms) between the reading end-dates of the two files (read on COME M – canal TMI i and on COME N – canal TMI j) is lower than OPS_DELAI_INTER_FIN_LEC seconds, then it is forbidden to continue (continuous reading) with the reading of COME N on canal i and of COME M on canal j."]

As one might expect, the present tense is used most often in all three corpora. No occurrences of the imperative were found, even in the exhortations of the requirements and the handbook, probably because it is too reader-oriented and also too direct to be used in such written texts, at least in French (it is obviously more common in oral instructions). Regarding the conditional verbs, it is not surprising that they are very rarely used in both the requirements and the handbook. More importantly, the future tense is used much more frequently in the requirements than in the other two corpora.

4.2. Qualitative analysis (of examples)

As a first step in the diagnosis, we focus on the description of some examples of phrases and sentences that do not follow the recommendations and try to understand why. We do not aim at criticizing these recommendations, but rather to see in which cases they seem relevant for our corpus and in which corpus they do not, and thus should be adapted.

4.2.1. Combinators

Some combinators are mandatory:

(1) "Le générateur de TCH vérifiera *que* la valeur du champ PHASE est comprise entre 0 *et* FREQ_DIV-1" ["The generator of TCH will check *that* the value of the field PHASE is between 0 *and* FREQ_DIV-1"]

In example 1, the subordinating conjunction *que* cannot be avoided, since it introduces the dependent clause[9], and the coordinating conjunction *et* is necessary to set the lower and higher limits of the interval.

Some combinators are not mandatory, but prevent repetitions and multiple sentences:

(2) "Les champs SM_ID *et* FM_ID seront extraits à partir de la BDS" ["Fields SM_ID *and* FM_ID will be extracted from the BDS"]

If the use of *et* were not allowed in example 2, two distinct sentences would be necessary ("Le champ SM_ID sera extrait à partir de la BDS" and "Le champ FM_ID sera extrait à partir de la BDS"). This would lead to longer and probably more confusing requirements: since the two sentences differ by only a single letter, the reader may not notice the difference and think it is a duplicated sentence.

However, longer sentences may become less readable:

(3) "Cette TC permet de passer contrôle thermique plate-forme en mode REDUCED, c'est-à-dire de sélectionner des seuils de régulation "larges" pour le contrôle thermique grossier (pour limiter la puissance consommée), *et* de modifier la valeur d'écrêtage de la puissance injectée pour le contrôle thermique fin" ["This TC makes it possible to switch the heat control of the platform to REDUCED mode, i.e. to select "broad" regulation thresholds for a coarse heat control (to limit the power consumed), *and* to change the cut-off value of the injected power for precise heat control"]

In example 3, it would probably have been better to clearly distinguish the two actions permitted by the TC – for example, with a bullet list.

Some combinators provide logical information that may help the reader to better understand the requirements:

(4) "pour n=2 la loi de la taille est respectée de fait *mais* le test 'FIFO vide' reste nécessaire" ["for n=2 the size rule is always respected, *but* the 'empty FIFO' test is still required"]

In example 4, the reader is certain that the test is necessary in all cases. Without the first main clause and the logical connector *mais*, he or she could have doubted it.

Nonetheless, in several cases, the use of a coordinator does not seem justified; in particular when two sentences are coordinated by *et*:

(5) "Le format des données de mesure angulaire et Doppler est conforme au standard CCSDS décrit dans le document DA9 *et* le schéma XML respecte le standard décrit dans DA11" ["The data format of the

[9] In French, the complementizer "que" is mandatory.

angular and Doppler measurement is in accordance with the CCSDS standard described in document DA9 *and* the XML schema complies with the standard described in DA11"]

(6) "Les demandes sont saisies sur le FOS *et* le logiciel ARPE gère les conflits entre les demandes Spot, Hélios et Pléïades" ["The requests are to be entered on the FOS *and* the ARPE software manages conflicts between the requests from Spot, Hélios and Pléïades"]

In examples 5 and 6, there is no apparent reason why separate sentences should not be used (parataxis); the use of the conjunction "and" is often not justified when it joins independent clauses without several common elements (subject, verb or object).

In some cases, problems arise because of the (absence of proper) coordinators:

(7) "Pour cela, on utilisera les données BDS (LENGTH et LOCATION_UNIT) de la table des OBCD (globaux) *ou* la description (LONGUEUR) des paramètres diagnostic déjà crées" ["For this, we will use the BDS data (LENGTH and LOCATION_UNIT) from the (global) OBCD table *or* the description (LONGUEUR) of the already created diagnostic parameters"]

In example 7 above, there are two possible solutions (alternative), but no explanation is given to the reader to tell him or her in which case(s) one of them should be preferred (or whether they are in fact identical).

(8) "Sur réception de cette TC, le LVC met à jour la table des surveillances standards de l'application destinataire *et* ré-initialise le compteur d'erreur (remise à 0) associé à cette surveillance" ["Upon reception of this TC, the LVC updates the table of standard surveillances of the destination application *and* resets the error counter associated to this surveillance"]

In example 8, we know that the LVC has to do two distinct operations, but it is not clear whether they are supposed to be done at the same time or one after the other.

(9) "(e.g : 2 *et* 10 *ou* 3 *et* 11)" ["e.g. 2 *and* 10 *or* 3 *and* 11)"]

In example 9, the priorities of the logical operators *et* and *ou* are not clear.

(10) "Cet ordre est rejeté *si* : ["This order is rejected *if*:"]
- le mode NORM automatique est actif ["the automatic NORM mode is active"]
- le satellite est en mode MAN ["the satellite is in MAN mode"]
- le satellite n'est pas en mode convergé (GAP ou SUP) ["the satellite is not in converged mode (GAO or SUP)"]
- un ordre MAN/CAP est déjà en attente d'exécution ["a MAN/CAP instruction is already waiting to be executed"]"

In example 10, the absence of coordinators between the items in the list is the source of uncertainty: is the order rejected if any of the following conditions is met (*or*), or only if they are all met (*and*)? Lists of this kind are very common in our corpus.

4.2.2. Pronouns

Some pronouns must be avoided, because otherwise the requirement is no longer autonomous:

(11) "*Il* calculera aussi, a 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" ["*It* will also calculate, at a frequency that can be parameterized (at monthly intervals), the average time for commissioning and will compare it to the maximum average in order to anticipate any problems"]

The requirement given in example 11 cannot be understood by itself, because the pronoun *il* refers to the subject defined in the previous requirement. This is also the case in another requirement, where a reference is made to a "previously stated rule", but without indication as to which rule is meant.

Some pronouns are mandatory:

(12) "Sur réception de cette TC, le LVC met à jour le paramètre *qui* donne la taille maximum d'un paquet TM de type dump" ["Upon reception of this TC, the LCV updates the parameter *that* gives the maximum size of a TM dump packet"]

Without the relative pronoun *qui*, it would not be possible to specify which parameter is referred to in example 12.

(13) "*Il* ne sera pas utile de vérifier ce paquet "vide"" ["*It* will not be necessary to check that "empty" packet"]

Impersonal pronouns like the one given in example 13 are widespread in our corpus and can hardly be avoided. They do not refer to another noun, cannot be replaced and are therefore not problematic.

Some pronouns are not mandatory, but prevent unnecessary repetitions of words:

(14) "La liste des TCD est définie en BDS. *Elle* est donnée ici à titre informatif:" ["The list of TCD is defined in BDS. *It* is given here for information:"]

Compare example 14 with the same sentences where the noun phrase is repeated in full: "La liste des TCD est définie en BDS. La liste des TCD est donnée ici à titre informatif:" ["The list of TCD is defined in BDS. The list of TCD is given here for information:"]

(15) "Le paquet ne sera généré que s'*il* est activé par le LVC" ["The packet will be generated only if *it* is activated by the LVC"]

Example 15 seems even less natural if rewritten without a pronoun: "Le paquet ne sera généré que si le paquet est activé par le LVC" ["The packet will be generated only if the packet is activated by the LVC"].

In short, personal pronouns should be used only if there is one and only one possible antecedent in the requirement. French demonstrative pronouns make it possible to avoid ambiguity between the subject and the object of a sentence:

(16) "Le générateur de TC ne rejettera pas la création du PARAM_ID diagnostic si *celui-ci* est déjà défini à bord" ["The TC generator will not reject the creation of the PARAM_ID diagnostic if *the latter* is already defined on board"]

In example 16, *celui-ci* refers to the closest noun and is therefore unambiguous (whereas *il* could have been ambiguous).

4.2.3. Pronoun *on*

In practice, thanks to its specific features (it can be used either as an indefinite pronoun or to replace any personal pronoun), *on* has many different purposes (and shortcomings) that we will try to classify.

On *does not refer to any particular person, and the clause in which it appears can be easily omitted:*

(17) "*On* considère que les autres TC sont associées à la famille F_AUTRES" ["We consider that the other TCs are associated with the family F_AUTRES"]

The main clause in example 17 is superfluous; the sentence could simply be written as "Par convention, les autres TC sont associées à la famille F_AUTRES" ["By convention, the other TCs are associated with the family F_AUTRES"].

(18) "*On* prendra comme valeur d'origine du syndrome la valeur 0XFFFF" ["We will take as initial value of the syndrome the value 0XFFFF"]

(19) "*On* recommande donc une durée d'observation de l'ordre de 25 à 100 sec" ["We therefore recommend a duration of observation between 25 and 100 seconds"]

Examples 18 and 19 could have been written respectively as "La valeur d'origine du syndrome sera 0XFFFF" ["The original value of the syndrome will be 0XFFFF"] and "La durée d'observation recommandée est donc de l'ordre de 25 à 100 secondes" ["The recommended duration of observation is thus between 25 and 100 seconds"]. These last three examples may be a hint that *on* is overused in the requirements.

On *is too vague:*

(20) "*On* vérifiera que les tables sont reçues en intégralité" ["One will check that the tables are fully received"]

(21) "*On* devra disposer de ces résultats en ligne depuis le tir" ["One will need these results on-line from the launch"]

The uses of *on* in examples 20 and 21 are problematic, since we have no indication about who is concerned by the instructions. The first one is a typical omission of the agent: no one (or nothing) is formally assigned to the task. In the second one, it is not clear at all who (or what) will need the results in question.

On *is deliberately general:*

(22) "*On* ne peut écrire et lire simultanément le même fichier" ["One cannot write and read the same file simultaneously"]
(23) "En conséquence *on* ne mixera jamais des clés C et M**"** ["Hence one will never mix the keys C and M"]

The vagueness of *on* has its benefits, and we may assume that, in examples 22 and 23, it is voluntarily used to make the rules more general: it plays the role of "everybody" or "nobody".

On *implicitly refers to the writers or to the readers:*

(24) "*On* ne listera ici que les TC associées aux familles F_EMETTEUR et F_MANOEUVRES" ["Here we will list only the TCs associated with the families F_EMETTEUR and F_MANOEUVRES"]
(25) "*On* se rapportera à DA14 pour la description de l'interface" ["See DA14 for the description of the interface"]

In example 24, *on* obviously refers to the writer, but is a convenient way to avoid the first person personal pronoun. Example 25 is an indication to the reader (who knows what to do in case he or she wants more information about the interface), but it is given in an indirect manner. This attempt to conceal the shifters (Fludernik 1991) makes the requirements less personal and more general.

On *to avoid the passive voice:*

(26) "*On* nommera cette durée DPDV_COEFF" ["We will call this duration DPDV_COEFF"]
(27) "*On* définira par ailleurs la notion […] par le triplet suivant" ["We will define the notion […] by the following triple"]

As we suspected, *on* is sometimes used when a passive voice would have been possible: "Cette durée sera nommée DPDV_COEFF" ["This duration will be called DPDV_COEFF"][10] for example 26, and "Par ailleurs, la notion sera définie par le triplet suivant" ["The notion […] will be defined by the following triple"] for example 27. These two examples are not really problematic, because the agent – supplemented by *on* – is not needed here, as in examples 17-19.

(28) "*On* contrôlera les bornes MIN et MAX de la FT BDS […] associée à ce paramètre" ["One will check the MIN and MAX boundaries of the FT BDS […] associated with this parameter"]
(29) "*On* initialisera le modèle quand la batterie est en charge complete" ["One will initialize the model when the battery is fully charged"]

In examples 28 and 29, however, the absence of the agent is problematic (as in examples 20 and 21), since we do not know who is in charge of performing these tasks (once again, *on* fills the role of subject of the sentence, but does not clearly refer to someone or something). If the sentences were written in the passive voice ("the boundaries will be checked", "the model will be initialized"), we would expect the agent to be introduced by the preposition "by" ("by the operator", "by the subsystem"). This omission is possible only when the agent is implicitly known to the reader, either because it is mentioned elsewhere in the requirement, or because it is part of the knowledge shared by the writers and the readers. However, while the responsibility of the agent is lower, the risk of confusion is greater.

All these examples illustrate the very different roles of *on* (some acceptable, like examples 22-25, others litigious, like examples 28 and 29), which may be hard to distinguish. In particular, it can be used to replace the passive voice – but this does not solve the problems related to the absence of the agent. In general, we can say

[10] Or, if we really want to use the active voice: "Le nom de cette durée sera DPDV_COEFF" ["The name of this duration will be DPDV_COEFF"].

that forbidding all kinds of passive sentences seems radical, because it introduces new difficulties (e.g. determining whether a past participle is used as an adjective) and because the potential problem of the agent is more specific (and should be addressed directly).

4.2.4. Future tense

We have shown that future verbs are quite common in the requirements we examined. The future may be used to state an injunction (remember that requirements are contractual obligations that must be complied with): without further clarification, it indicates that the task will have to be done (after the time at which the requirement was written), one or several times. See examples 30 and 31:

(30) "Le vidage des tables *sera* contrôlé par le CCC" ["The emptying of the tables *will* be checked by the CCC"]

(31) "On *affichera* donc trois valeurs" ["We *will* thus show three values"]

Just like the present tense, the future can express different linguistic modalities. They can be specified or strengthened by modal verbs:

(32) "Le plan TC *devra* respecter les contraintes décrites dans DR20" ["The TC plan *will have to* respect the constraints described in DR20"]

In example 32, the modal verb "devoir" insists on the directive modality, but is not essential, since (given the context) the sentence would have a similar meaning if it were not present. In fact, many examples of the same kinds of injunction expressed with or without a modal verb can be found in our corpus (e.g. "pourra" *vs.* "devra pouvoir" ["will (have to) be able to"]).[11]

In the same way, adverbs are sometimes useful:

(33) "Les champs DONNEE *seront systématiquement* initialisés à zéro" ["The fields DONNEE *will be systematically* initialized to zero"]

And the absence of obligation can also be expressed:

(34) "Il *ne sera pas utile* de vérifier ce paquet "vide"" ["It *will not be necessary* to check this "empty" packet"]

Therefore, we may wonder why writers sometimes use the present and sometimes the future tense, when both seem possible.[12] We fear that this hesitation (that we interpret as a consequence of the tension between the time when the requirement was written, and the time when the system is implemented, that is itself specific to this genre of technical texts) might be confusing, especially when it appears inside the same requirement. The specifications would probably be easier to understand if the tenses were chosen consistently: future verbs for actions that remain to be completed and present verbs in all other cases (e.g. to express universal facts).

4.2.5. Towards new rules

This simple analysis of two rules from INCOSE shows us that they are probably too restrictive and that in several cases, they cannot be strictly applied. As one would suspect, conjunctions and pronouns remain frequent in technical texts and it would be too constraining (and therefore counter-productive) to oblige the engineers to avoid them by all means.

Nevertheless, they address concrete problems such as sentence complexity or anaphora resolution and should not be ignored, but rather refined:

[11] In the sentences using a present verb, however, it is not always clear. In the following example: "L'opérateur […] *peut*, à tout moment, se connecter/déconnecter du système d'exploitation" ["The operator *can*, at any time, connect to or disconnect from the operating system"], is this already the case, or is it a condition to be met by the system – and therefore an instruction?

[12] Example: "Le CCC met à disposition du COO :" ["The CCC provides the COO with:"].

- Regarding conjunctions, most of them are useful (if not unavoidable[13]) if properly used. We identified only one case where a conjunction ("and") was totally unnecessary and should preferably be avoided: when it links two independent clauses. Indeed, such clauses should better be part of distinct requirements, in order to respect the "singularity" principle advocated by INCOSE and other guides. However, it can be useful to join these two clauses if they share two of the three following elements: subject, verb and complement. This similarity tend to indicate that they are probably very close to one another (and that consequently they can be treated together), and the conjunction "and" can avoid an unwelcome repetition. (This conclusion could partially be criticized because two different verbs may indicate two different processes; this issue should be investigated more in detail.) Most of the problems arising from conjunctions are actually mostly due to their absence: for instance, writers should always precise if only one or all the conditions are needed.
- Regarding pronouns, the requirements would completely lose their naturalness if they were forbidden. A better rule should state that a pronoun can be used only if it has one (and only one) possible antecedent within the requirement: if it has more than one possible referent, then it is ambiguous; if it has none, then the requirement is no longer autonomous. Besides, this rule should also cover determiners, which brings similar problems.

Furthermore, rules can also be drawn from some of the characteristics observed in the corpus. For example, a rule could impose a more consistent use of grammatical tenses: the future for actions that will be performed later, and the present for affirmations that are already true at the time of the writing.

4.3. From single words to structures

In the previous sections, we explored only how single words function. Obviously, the main characteristics of CNLs concern structures rather than single words. In our viewpoint, in most cases, single words may be considered as pivots of more complex structures. For example, the pronoun *on* occurs very often with future tense. Then, in our corpus, it is the structure using *on* as subject with a verb on future that constitutes a recurrent pattern. This combination seems to be a characteristic of the genre we are investigating.

With such a point of view, in a recent study (Warnier and Condamines 2015), we used the results produced by a "candidate-terms" extractor as the first step to access the recurrent syntactic structures in which they occur. Using the results obtained in two requirement corpora, we first filtered the list of candidate-terms by eliminating the terms belonging to the space domain. We then considered that most of the remaining candidate-terms were specific not to the domain but more probably to the genre of requirements at CNES. These "terms" were used to access to the specific structures of the requirement genre.

5 Conclusion

This paper aimed to show that it is possible to bridge two kinds of norms, one laid down by an external body (*normalisation*) and one emerging from a corpus of requirements (*normaison*) in order to build a CNL adapted to requirements writing. At this stage in the study, the corpus comprises the requirements of a space project (Pleiades), and the recommendations come from two different existing CNLs (INCOSE, ASD STE). The method used combines both corpus-based and corpus driven-approaches. The corpus-based approach is based on some of the (direct or indirect) recommendations from the CNLs that we have interpreted in terms of linguistic phenomena. The corpus-driven approach reveals what characteristics seem specific to the corpus. The results obtained by exploring the corpus allowed us to identify new hypotheses and then to verify them on the corpus.

These results of our initial evaluation are quite encouraging with respect to the suitability of our method. However, many analyses still remain to be done.

First, it is necessary to increase the size of the corpus in order to ascertain whether the characteristics of the requirements observed for Pleiades also feature in other specifications from the CNES. The specific topic of a project is unlikely to play an important role, at least within the same firm, in the characteristics of a "requirements" corpus. Probably, only part of the lexicon will vary across projects; the main part is not expected

[13] For instance, the sentence "The system will provide black and white pictures" is not equivalent to "The system will provide black pictures" + "The system will provide white pictures".

to vary significantly. It is however necessary to validate this hypothesis of the existence of only one textual genre for requirements writing (at CNES and possibly in other firms). We have therefore built a new corpus of requirements, extracted from the specifications of another (smaller) CNES project. From a practical point of view, we do not know to what extent it will be necessary to build new CNLs for other firms and/or for other products. We consider, however, that our method should be applicable to other domains or fields and even to other languages.

Second, we need to better analyze how our method may be completed by data-mining tools and tools focusing on the identification of the grammar of genre. AntConc (Anthony 2005) and its statistical functionality, and SubCAT, to compare the text to a reference corpus, are possible candidates. We have used SDMC (Quiniou et al. 2012), oriented towards the characterization of the grammar of genre, and the results seem to be useful in the aim to confront the recurrent patterns obtained with the recommendations of the CNLs (Warnier et Condamines 2015).

Third, we plan to analyze other characteristics present in CNLs. For instance, we intend to study nominalizations. It is well known that they are more abundant in specialized corpora than in newspapers or literature, both in French (Condamines and Bourigault 1999) and in English (Nishina 2007). If this is also the case within requirements, it could be interesting to see why and whether this over-abundance is linked to the possibility of being more evasive with nominalizations than with verbs (nominalizations are more polysemic than verbs, and it is not necessary to specify all the arguments). We also intend to propose a finer-grained analysis of the passive voice and of *if*-conditionals, and we assume that the way negation is used in the requirements could be of interest. We will also look for definitions and examples in the requirements, as they are (in theory) to be proscribed. One of the main issues will be to examine how to consider CNL recommendations that are not really explicit from a linguistic point of view. It is not easy to translate all the recommendations into linguistic forms. For example, how should one interpret the rule that "a requirement statement should address a single consideration" or that "a requirement statement should be expressed at a level of detail appropriate to its level of abstraction" (two rules from INCOSE)?

Finally, the utility and efficiency of the CNL (when completed) will have to be evaluated "in the field": the simplest way to decide whether it is useful to the engineers will be to ask them their opinion and, if possible, to verify if the rules are really applied. In particular, if they are implemented in an automatic tool (a longer-term project), it will be very interesting to review the suggestions proposed to the writers to see which ones were accepted or not, and to consider their justifications.

Acknowledgments This study is carried out as part of a PhD thesis granted by the CNES and the Regional Council Midi-Pyrénées. We would like to thank the CNES for their active cooperation as well as for providing us with the requirements corpus. We are also very grateful to the anonymous reviewers of this special issue and to those of the Fourth Workshop on Controlled Natural Language (CNL 2014), for all their relevant comments, suggestions and references.

References

AeroSpace and Defence Industries Association of Europe. (2007). Simplified Technical English. Specification ASD-STE100. International specification for the preparation of maintenance documentation in a controlled language. Issue 4.

Anthony, L. (2005). AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom (pp. 729–737). Presented at the Professional Communication Conference, 2005. IPCC 2005. Proceedings. International. doi:10.1109/IPCC.2005.1494244

Austin, J. L. (1975). *How to Do Things with Words*. Oxford University Press.

Bakhtin, M. (2004). *Speech Genres and Other Late Essays*. University of Texas Press Slavic Series.

Barcellini, F., Albert, C., Grosse, C., and Saint-Dizier, P. (2012). Risk Analysis and Prevention: LELIE, a Tool dedicated to Procedure and Requirement Authoring. In *LREC*, 698-705.

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

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Biber, D. (2006). *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins Publishing.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International journal of corpus linguistics*, *14*(3), 275–311.

Bouquet, S. (2007). Contribution à une linguistique néo-saussurienne des genres de la parole (1) : une grammaire du morphème on. *Linx. Revue des linguistes de l'université Paris X Nanterre*, (56), 143–156. doi:10.4000/linx.376

Bruce, B., Rubin A., Starr K.S. (1981). Why readability formulas fail. In *IEEE Transactions on Professional Communication*, PC-24, pp. 50-52.

Cabré, M. T. (1999). *Terminology: Theory, Methods, and Applications*. John Benjamins Publishing.

Carlson, N. and Laplante, P. (2013), The NASA automated requirements measurement tool: a reconstruction. In *Innovations Syst Softw Eng*, 10(2), 77–91.

Chervak, S., Drury, C. G., & Ouellette, J. P. (1996). Field evaluation of simplified english for aircraft workcards. In *Proceedings of the 10th FAA/AAM Meeting on Human Factors in Aviation Maintenance and Inspection*.

Clark, P., Murray, W. R., Harrison, P., & Thompson, J. (2010). Naturalness vs. Predictability: A Key Debate in Controlled Languages. In N. E. Fuchs (Ed.), *Controlled Natural Language* (pp. 65–81). Springer Berlin Heidelberg.

Condamines, A. (1995). Terminology: New needs, new perspectives. *Terminology*, *2*(2), 219–238. doi:10.1075/term.2.2.03con

Condamines, A. & Bourigault, D. (1999). Alternance nom/verbe : explorations en corpus spécialisés. In Victorri, B. & François, J. (Ed.), *Sémantique du lexique verbal, Actes de l'atelier de Caen, 22-23 janvier 1999*, Cahiers de l'Elsap, 41-48.

Condamines, A. (2010). Variations in terminology: Application to the management of risks related to language use in the workplace. *Terminology*, *16*(1), 30–50. doi:10.1075/term.16.1.02con

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

Firth, J. R. (1957). *Papers in linguistics, 1934-1951*. Oxford University Press.

Flesch, R. (1948). A New Readability Yardstick. *Journal of Applied Psychology*, 32(3), 221-233.

Fludernik, M. (1991). Shifters and deixis: Some reflections on Jakobson, Jespersen, and reference. *Semiotica*, *86*(3-4), 193–230. doi:10.1515/semi.1991.86.3-4.193

Freixa, J. (2006). Causes of denominative variation in terminology: A typology proposal. *Terminology*, *12*(1), 51–77. doi:10.1075/term.12.1.04fre

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

Gilliland, J. (1972) *Readability*. London: University of London Press Ltd.

Guespin, L. (1993). Normaliser ou standardiser? *Le Langage et l'homme*, *28*(4), 213–222.

Hymes, D. (1967). Models of the Interaction of Language and Social Setting. *Journal of Social Issues*, *23*(2), 8–28. doi:10.1111/j.1540-4560.1967.tb00572.x

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

Jakobson, R. (1960). Linguistics and Poetics. In T. Sebeok (Ed.), *Style in Language* (M.I.T. Press., pp. 350–353). Cambridge.

Kittredge, R., & Lehrberger, J. (1982). *Sublanguage: Studies of Language in Restricted Semantic Domains*. Walter de Gruyter.

Klare, G. R. (1976). A second look at the validity of readability formulas. *Journal of Reading Behavior*, 8, 129-152.

Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, *40*(1), 121–170. doi:10.1162/COLI_a_00168

Kurzon, D. (1997). "Legal language": varieties, genres, registers, discourses. *International Journal of Applied Linguistics*, *7*(2), 119–139. doi:10.1111/j.1473-4192.1997.tb00111.x

Le Querler, N. (2004). Les modalités en français. *Revue belge de philologie et d'histoire*, *82*(3), 643–656. doi:10.3406/rbph.2004.4850

Lopez, S., Condamines, A., Josselin-Leray, A., O'Donoghue, M., & Salmon, R. (2013). Linguistic Analysis of English Phraseology and Plain Language in Air-Ground Communication. *Journal of Air Transport Studies*, *4*(1), 44–60.

Malrieu, D. (2007). Contribution à une linguistique néo-saussurienne des genres de la parole (2) : analyse des valeurs d'indexicalité interlocutoire de on selon les genres textuels. *Linx. Revue des linguistes de l'université Paris X Nanterre*, (56), 157–178. doi:10.4000/linx.377

McEnery, T. & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Meyer, B. (1985). On Formalism in Specifications. In *IEEE Softw.*, 2(1), 6–26.

Nishina, Y. (2007). A Corpus-Driven Approach to Genre Analysis: The Reinvestigation of Academic, Newspaper and Literary Texts. *Empirical Language Research*, *1*.

O'Brien, S. (2003). Controlling Controlled English. An analysis of several controlled language rule sets. *Proceedings of EAMT-CLAW*, *3*, 105–114.

Pace, G. J., & Rosner, M. (2010). A Controlled Language for the Specification of Contracts. In N. E. Fuchs (Ed.), *Controlled Natural Language* (pp. 226–245). Springer Berlin Heidelberg.

Paumier, S. (2011). Unitex - Manuel d'utilisation.

Pearson, J. (1998). *Terms in Context*. John Benjamins Publishing.

Poudat, C. (2003). Characterization of French linguistic research articles using morphosyntactic variables. *Academic discourse. Multidisciplinary approaches*, pp. 77–95. Oslo.

Quiniou, S., Cellier, P., Charnois, T., & Legallois, D. (2012). What about Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics? In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 166–177). Springer Berlin Heidelberg.

Shubert, S. K., Spyridakis, J. H., Holmback, H. K., & Coney, M. B. (1995). The comprehensibility of simplified English in procedures. *Journal of technical writing and communication*, *25*(4), 347-369.

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* (pp. 131–139). Stroudsburg, PA, USA: Association for Computational Linguistics.

Stage, L. (2002). Les modalités épistémique et déontique dans les énoncés au futur (simple et composé). *Revue romane*, *37*(1).

Stewart, K. M. (1998). Effect of AECMA simplified English on the comprehension of aircraft maintenance procedures by non-native English speakers. Master's thesis, University of British Columbia.

Swales, J. (2004). *Research Genres: Exploration and Applications*. Cambridge University Press.

Temmerman, R. (2000). *Towards New Ways of Terminology Description: The Sociocognitive-approach*. John Benjamins Publishing.

Temnikova, I. (2012). *Text Complexity and Text Simplification in the Crisis Management domain* (PhD thesis). University of Wolverhampton, UK.

Temnikova, I., Baumgartner Jr, W. A., Hailu, N. D., Nikolova, I., McEnery, T., Kilgarriff, A., Angelova, G. & 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)*.

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.

Warnier, M., & Condamines, A. (2015). A Methodology for Identifying Terms and Patterns Specific to Requirements as a Textual Genre Using Automated Tools. In *Proceedings of the International Conference "Terminology and Artificial Intelligence*. 4-6 November, Granada, Spain, 183-190.

Wüster, E. (1968). *The Machine Tool: an Interlingual Dictionary of Basic Concepts*. Technical Press.

Wyner, A., Angelov, K., Barzdins, G., Damljanovic, D., Davis, B., Fuchs, N., et al. (2010). On Controlled Natural Languages: Properties and Prospects. In N. E. Fuchs (Ed.), *Controlled Natural Language* (pp. 281–289). Springer Berlin Heidelberg.

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