

Linguistic Analysis of Requirements of a Space Project and Their Conformity with the Recommendations Proposed by a Controlled Natural Language

Anne Condamines and Maxime Warnier
{anne.condamines,maxime.warnier}@univ-tlse2.fr

CNRS – CLLE-ERSS

Fourth Workshop on Controlled Natural Language
Galway, 20 August 2014



Contents

Introduction

Aim

Current situation

Hypothesis

Methodology

First Results

Conclusion

General aim: improve the *writing* of requirements* at the CNES (upstream work: change the way engineers write them)

Issue: requirements are critical (*contractual* obligations); divergent interpretations may lead to

- ▶ additional costs
- ▶ delays
- ▶ litigation

especially for large-scale projects

* "*conditions or capabilities that must be met or possessed by a system or system component to satisfy a contract, standard, specification, or other formally imposed document*" [IEEE 1990]

Requirements are part of *specifications*, which may also include other comments or examples, and each requirement should be autonomous

Currently, CNES engineers do *not* use a Controlled Language (only requirement management tools)

Natural Language

- ▶ is the most expressive way to specify needs
- ▶ is easy to use among stakeholders
- ▶ can never be totally avoided
- ▶ is often vague and ambiguous

Goal: a Controlled Natural Language that still looks natural, but limits ambiguity as much as possible

Hypothesis: even if engineers are not obliged to follow any CNL, they could be (indirectly) *influenced* by the existing guidelines – because they are experienced and used to read other requirements

Our first contribution is to try to identify *linguistic regularities* in their written productions and to prove the existence of a *textual genre**

* "*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]

Unlike that of *sublanguage*, this notion comes from a sociolinguistic approach

In the French school of Rouen, these spontaneous regularities have been called *normaison*, as opposed to *normalisation* (linguistic norms imposed by an organism)

This *normaison* could be a useful starting point for a *normalisation*

Contents

Introduction

Methodology

Approach

Corpus

Controlled Natural Language

First Results

Conclusion

Several existing CNL for writing requirements are not close enough to established practices and are

- ▶ sometimes too constraining (inapplicable)
- ▶ sometimes insufficiently so (problems not solved)
- ▶ not always rigorous

They are, in general, designed by *domain* experts – not by language experts; their experience is valuable, and so should be our knowledge of the language and our analysis based on *actual data* (corpus linguistics)

Methodology: *comparison* between rules from a CNL (INCOSE) and requirements from the CNES to see whether the latter are in accordance with the former (if not, why?)

(Remember that engineers were not asked to follow those rules!)

Reference corpus composed of 1,142 requirements (nearly 53,000 words) in French, extracted from a subset of the specifications of the project Pleiades (two very high resolution Earth observation satellites launched in 2011 and 2012)

Two **other corpora** of the same size:

- ▶ a handbook written by experts from the CNES, intended for semi-experts, about techniques used for operating spacecrafts
- ▶ some articles from the newspaper *Le Monde*

Guide for Writing Requirements proposed by INCOSE 2011 (International Council on Systems Engineering):

- ▶ widely based on former standards and guidelines (rules are often similar among them)
- ▶ quite general (not limited to a single discipline)
- ▶ facilitates human-to-human communication

It can be called a *Controlled Natural Language* according to the definition given by Kuhn 2014:

- ▶ it is a constructed language
- ▶ it has one base language (English)
- ▶ it sets constraints on the vocabulary, the syntax and the semantics
- ▶ requirements are still understandable by English speakers

Two rules selected because they are not highly language-dependent and can be verified in a semi-automated manner:

1. Singularity/Propositionals: **Avoid combinators**
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.
2. Completeness/Pronouns: **Repeat nouns in full instead of using pronouns to refer to nouns in other requirement statements**
Pronouns are words such as 'it', 'this', 'that', 'he', 'she', 'they', 'them'. When writing stories, they (sic.) are a useful device for avoiding the repetition of words; but when writing requirements, pronouns should be avoided, and the proper nouns repeated where necessary.

These rules

- ▶ lack of precision (only give examples of combinators and pronouns, do not specify when it is 'necessary')
- ▶ are very restrictive (avoid all combinators and pronouns?)
- ▶ have evasive justifications
- ▶ are not consistent (examples contradict them)

Contents

Introduction

Methodology

First Results

Quantitative analysis

Analysis of examples

Conclusion

Number of conjunctions and pronouns:

	Coordinators	Subordinators	Conjunctions	Pronouns
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%)

- ▶ Both are numerous in the three corpora
- ▶ But they are much less frequent in requirements (cf. hypothesis: spontaneous regularities)

Length of sentences:

	No. sentences	No. long sentences	Average sentence length
Requirements	4859	350 (7.20%)	11
Handbook	3456	591 (17.10%)	15
Newspaper	2201	839 (38.10%)	24

- ▶ Long sentences (containing more than 25 words) are numerous in the three corpora
- ▶ But they are much less frequent in requirements (cf. hypothesis: spontaneous regularities)
- ▶ Longest sentence in requirements: *"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."*

Combinators

Combinators

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`*

Combinators

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*

Combinators

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*

Combinators

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*

- ▶ Some conjunctions are *mandatory* and some are *useful* because they avoid repetitions or provide logical information
- ▶ The conjunction "and" is not justified when it joins independent clauses without a common element (subject, verb, complement)
- ▶ In other cases, it can often be replaced by a bullet list

Pronouns

Pronouns

Il ne sera pas utile de vérifier ce paquet "vide"

It won't be necessary to check that "empty" packet

Pronouns

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:*

Pronouns

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.*

Pronouns

Il calculera aussi, a une fréquence paramétrable (ordre de grandeur 1 mois), la moyenne de mise en œuvre

It will also calculate, at a frequency that can be parameterized (at monthly intervals), the average time for commissioning

- ▶ Some pronouns are *mandatory* and some are *useful* because they avoid repetitions
- ▶ Personal pronouns should be used if there is one and only one possible antecedent in the requirement

Contents

Introduction

Methodology

First Results

Conclusion

Conclusions:

- ▶ Analysis of genuine requirements of a space project
- ▶ Albeit not written according to a CNL, they show interesting differences compared to other texts (fewer pronouns and conjunctions, shorter sentences)
- ▶ The rules from the CNL we considered (INCOSE) are useful guidelines, but are inapplicable as is and must be refined

Future work:

- ▶ Inventory existing rules in French CNL
- ▶ Automatically compare them with different corpora
- ▶ Propose a set of rules more consistent and closer to real practice

References

- ▶ ASD: Simplified Technical English. International specification for the preparation of maintenance documentation in a controlled language (Issue 4) (2007)
- ▶ Bhatia, V.K.: Analysing genre: Language use in professional settings. Longman, London (1993)
- ▶ Bünzli, A., Höfler, S.: Controlling Ambiguities in Legislative Language. In: Rosner, M. and Fuchs, N. (eds.) CNL 2010. pp. 21–42. Springer, Berlin Heidelberg (2012)
- ▶ Condamines, A.: Variations in Terminology. Application to the Management of Risks Related to Language Use in the Workplace. Terminology. 16, 30–50 (2010)
- ▶ Fuchs, N., Schwitter, R.: Controlled English for Requirements Specifications. IEEE Computer Special Issue on Interactive Natural Language Processing (1996)
- ▶ International Council on Systems Engineering: Guide for Writing Requirements (2011)
- ▶ Kuhn, T.: A Survey and Classification of Controlled Natural Languages. Computational Linguistics. 40, 121–170 (2014)
- ▶ O'Brien, S.: Controlling Controlled English. An Analysis of Several Controlled Language Rule Sets. Proceedings of EAMT-CLAW. pp. 105–114 (2003)