How can corpus linguistics help improve requirements writing? Specifications of a space project as a case study

Maxime Warnier

▶ **To cite this version:**

Maxime Warnier. How can corpus linguistics help improve requirements writing? Specifications of a space project as a case study. 23rd IEEE International Requirements Engineering Conference (RE'15), Aug 2015, Ottawa, Canada. pp.388 - 392, 10.1109/RE.2015.7320456 . hal-01422897

HAL Id: hal-01422897
https://hal.science/hal-01422897

Submitted on 27 Dec 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

How Can Corpus Linguistics Help Improve Requirements Writing? Specifications of a Space Project as a Case Study

Maxime Warnier
CLLE-ERSS (UMR5263 : CNRS & Université Toulouse – Jean Jaurès), CNES
Toulouse, France
maxime.warnier@univ-tlse2.fr

***Abstract*—The specific purpose of this doctoral research is to improve the writing of requirements at the French Space Agency (CNES) by proposing a set of linguistic rules – referred to as a Controlled Natural Language (CNL) – that engineers should follow when writing out specifications in French. CNLs for technical writing do already exist, but if they are reviewed from a linguistic point of view, they are found unsatisfactory and too constraining, because some of the rules they impose lack relevance or are not compatible with the way engineers actually specify large-scale systems. In this research abstract, we will present a methodology based on corpus analysis aimed at improving existing rules and suggesting new ones that are inspired by existing data. We will also consider requirements extracted from specifications written at CNES to demonstrate its feasibility.**

***Index Terms*—requirements writing, technical writing, controlled natural language, corpus linguistics, textual genre.**

I. Title of the Research

"Comparative linguistic analysis of technical writing guides and real specifications of space systems at CNES to improve the writing and understanding of requirements."

("Analyse linguistique comparée des guides de rédaction technique et des usages réels dans les spécifications de systèmes spatiaux au CNES en vue d'améliorer la rédaction et la compréhension des exigences".)

II. Research Problem

As the overriding importance of requirements is now widely acknowledged, advanced Requirements Engineering (RE) methods and tools are more and more in use in companies and institutions [1]. One of the most critical steps of RE activities – albeit sometimes underestimated – is the writing of specifications, i.e. documents containing all the requirements that will serve as a basis for the implementation of the system (and thus constitute the raw material of RE).

Indeed, such specifications, being both written and read by human beings, may in some cases and under certain circumstances lead to misunderstandings – which, in turn, may cause delays, supplementary costs, litigation between stakeholders (if the requirements are part of a contract), or even accidents.

Problems of this kind arise when the language used to communicate information is not perfectly univocal, as is the case with natural languages (languages used in everyday life, such as English, Spanish or French). The undesirable properties of documents written in unrestricted natural language are well-known [2]: ambiguity (lexical, syntactic or referential) [3], vagueness [4], incompleteness, and so on. As a consequence, many solutions have been proposed to improve the quality of these documents, such as natural language processing tools for semi-automatic verification [5], more formal languages (some of them very close to mathematical expressions and logic notations [6], and hence unambiguous), and linguistic rules to avoid problematic words, phrases and sentences. While the first solution is merely an aid offered to the users during or after the writing (downstream work) to ensure their production complies with predefined rules (that they may or not know), the last two require them to learn these languages or rules, known as Controlled Natural Languages (CNLs) [7], before they can actually start writing the specifications (upstream work); they are consequently more demanding – and this is especially the case when the rules and languages in question do not look "natural" at all. (See [8] for a distinction between "naturalist" vs. "formalist" CNLs, and [7] for a definition of "naturalness" in CNLs.) Naturally, all of the above-mentioned options have their own benefits and tradeoffs; and, to a certain extent, they can be combined. In any case, natural language remains easier to use among stakeholders and, due to its expressiveness, unavoidable, at least in the early steps of the projects.

III. Aim of the Research

The research originated from a request by the Quality Assurance sub-department of CNES (Centre National d'Études Spatiales, National Center for Space Studies), the French Space Agency. Engineers at CNES are in charge of designing innovative space systems (such as the probe Rosetta, which landed on a comet in 2014) and work with other companies as well as with scientific and military partners. In this context (large-scale projects involving many people over several years), the good comprehension of the requirements (which are written in natural language) is particularly crucial – and so is the quality of the drafting.

At the present time, although the use of software for requirement management and traceability (Rational DOORS,

978-1-4673-6905-3/15 © 2015 IEEE

RE 2015, Ottawa, ON, Canada
Doctoral Symposium

Accepted for publication by IEEE. © 2015 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/

Reqtify) by the engineers is systematic, they are not asked to follow any linguistic norm. Since they do not wish to use only formal notations in specifications, our objective is to propose a set of rules in the form of a (naturalist) CNL for requirements writing. In order for these rules to be actually applied, they need to remain close enough to what engineers are already used to read and write; if not, they will probably see them as an excessive constraint and simply ignore them.

To achieve this, we assume that it is important for us to rely on previous work: this means we would like to take into account already existing CNLs and other technical writing guides, but also to verify our hypotheses on genuine texts to give the rules a real foundation, not only intuitions. This methodological basis, which is an original aspect of our contribution, will be presented in greater detail in the following sections.

IV. Theoretical Considerations

Because so far no norms have been imposed to the engineers at CNES, we are free to propose our own, new CNL. Nonetheless, as already mentioned, we want the rules that compose this CNL to be as close as possible to the way engineers currently write requirements. This leads us to consider two kinds of documents as a starting point: existing CNLs on the one hand and authentic specifications (written for older projects) on the other. The former are representative of a prescriptive vision of technical writing, whereas the latter allow us to propose a descriptive point of view of requirements writing at CNES. Although the conclusions we can draw from these two document types are sometimes rather different, we believe that it would be interesting to combine them.

We are reviewing two distinct CNLs for technical writing. The first one, proposed by the AeroSpace and Defence Industries Association of Europe, is called *Simplified Technical English* [9] (from now on, ASD-STE) and is a well-known reference for maintenance documentation. The second one is the *Guide for Writing Requirements* [10] by the International Council on Systems Engineering (from now on, INCOSE), whose goal is "to draw together advice from a variety of existing standards into a single, comprehensive set of rules and objectives". Since INCOSE is presented as the state of the art of guidelines for requirements writing, some (but not all) of the rules that it defines (or at least similar ones) can also be found in older guides, including ASD-STE. The rules that we decided to focus on are explained in section V, but we can already state that, like most CNLs intended for firms, ASD-STE and INCOSE were designed by domain experts (in this case, engineers), not by language experts (linguists). As a result, some of their prescriptions are not totally appropriate, as is shown by the results we obtained from the comparison with the specifications (given in section VI). Nevertheless, we are convinced that their experience and knowledge of the field are of valuable help and cannot be neglected; most of the rules were proposed to avoid specific problems related to natural language, but are not always properly expressed.

By contrast, the specifications on which our analyses are conducted are supposed to be representative of the way engineers at CNES actually write requirements. Given this assumption, we would like to characterize this particular style of writing (in order to propose more realistic rules). That is, we need to identify the typical words, patterns or syntactic structures used to express a specific piece of information.

At first, this idea might seem to contradict the fact that no instructions are given to the engineers on how to write the requirements: in other words, there could be just as many different styles as there are people in charge of writing requirements. Still, we believe that spontaneous regularities are likely to arise in practice, because these people form a professional community and this task is a recurring, well-identified communicative situation. Here we refer to the notion of *textual genre*, defined by Bhatia 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" [11]. (Note that it is close to that of *sublanguage*, defined by Somers 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" [12], but the latter comes from a distributional approach.) In more concrete terms, this means the writers are probably influenced by their frequent interactions with their colleagues, by their own experience of reading and writing specifications, and maybe even – indirectly – by some rules that they may have read about or heard of. Therefore, on a more theoretical level, if we are able to spot linguistic regularities in the specifications we analyze (and to evidence that they are more frequent than in other text types), this could prove the existence of a textual genre of requirements written in French at CNES, with its own grammar. (A generalization to requirements written in other companies or in another language, for instance, would be a further step.) Corpus linguistics methods and tools seem to be perfectly suitable for these tasks.

Lastly, the two sides of requirements writing we want to build upon – external standards and spontaneous regularities – can be linked to the notions of "normalisation" and "normaison", proposed by Guespin for terminology [13] and later used in the field of socioterminology [14]. Both refer to a linguistic norm that speakers must follow if they want to be identified as members of the community [15]; the key difference is that "normalisation" is prescriptive and consciously defined, whereas "normaison" is descriptive and unconscious – and thus very close to a textual genre. So, we can also claim that we wish to propose a "normalisation" inspired by a "normaison".

V. Methodological Considerations

As discussed in the preceding sections, we want to identify typical linguistic features in the documents we were provided by CNES using well-established corpus linguistics methods. We can distinguish between two different approaches (although in practice they are not radically opposed and are often complementary): the corpus-based approach and the corpus-driven approach, in Tognini-Bonelli's words [16]. According to Biber, *corpus-based* research "assumes the validity of linguistic forms and structures derived from linguistic theory. The primary goal of research is to analyse the systematic patterns of variation and

use for those pre-defined linguistic features", and *corpus-driven* research "is more inductive, so that the linguistic constructs themselves emerge from analysis of a corpus" [17].

Following the corpus-based approach, we intend to first build hypotheses (mainly thanks to the recommendations proposed by the CNLs described in section IV), and then verify them on our corpus of requirements. Following the corpus-driven perspective, we basically rely on tools to make emerge specificities from the corpus that can be interpreted in terms of a grammar of genre.

Some of these hypotheses are directly based on explicit rules in one of the CNLs (depending on how they are formulated, it may be necessary to slightly adapt them):

- One rule from ASD-STE imposes a limited number of words per sentence (20 for procedural text or 25 for descriptive text). Other rules specify how to count words.
- One rule from INCOSE ("Singularity/Propositionals") recommends to "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." However, some of these so-called combinators are present in the examples of "acceptable" requirements (e.g. "The 'control side lamps' function shall illuminate the side lamps *while* any combination of the following lights is illuminated: […]").
- One rule from INCOSE ("Completeness/Pronouns") asks the writer to "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". But in the only example of unacceptable requirement, the possible ambiguity is due to a determiner, not a pronoun ("The controller shall send the driver *his* itinary [sic] for the day").

Hence, our hypotheses are: shorter sentences, fewer conjunctions, fewer pronouns. Although the rules we consider were proposed for English, we assume that these three phenomena are valid for French as well because they are not highly language-dependent.

Some of our hypotheses are extrapolated from other rules found in CNLs. For example, many of them forbid the passive voice [18] – or impose the active voice, as is the case in both ASD-STE and INCOSE –, because (among other reasons) it allows omission of the agent. But we think that writers may be tempted to find other ways to avoid specifying the agent; in French, one of these ways could be the third-person singular subject pronoun "on" (which can be used instead of any other subject pronoun [19, 20]). Hence, our hypothesis is: more pronouns "on".

Finally, some salient features of the corpus are revealed by statistical tools. For instance, it is remarkable that numerous requirements are expressed with the future tense. This can be explained in part because the system described in the requirements and its components do not yet exist, but probably also because the future is seen as a "less direct form of instruction", as stated in ASD-STE, which prefers the imperative and adds that such forms "leave confusion as to whether something: must be done, or is already done, or must be done in the future by someone else". Moreover, in instructions written with the imperative, the agent is clearly identified (i.e. the reader), but this is not necessarily the case in sentences written with the future tense; for this reason, we believe it would make sense to analyze this phenomenon in conjunction with the pronoun "on" (see above). Hence, our hypothesis is: more sentences with the future tense and the pronoun "on".

To test our hypotheses and propose a "diagnosis", we will use three comparable corpora (same language: French; same number of words: 53,000). The reference corpus is composed of 1,142 requirements extracted from a subset of the specifications of the project Pléiades (two very-high-resolution Earth observation satellites launched in 2011 and 2012).[1] The two comparison corpora are taken from:

- a handbook about techniques and technologies used for building and operating spacecraft (it is written by experts from CNES and intended to semi-experts);
- articles from the French national newspaper *Le Monde*.

The former should represent a technical text in the same domain, whereas the latter should constitute a "generic" corpus, and both should help us situate the reference corpus.

As a conclusion to this section, the general methodology is illustrated by Fig. 1.

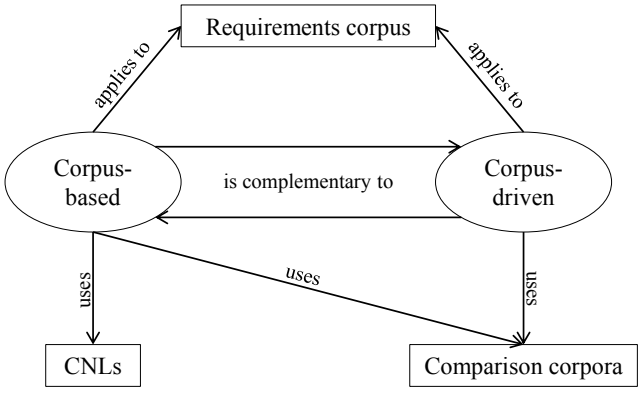

Fig. 1. Methodology.

VI. Results

In this section, we would like to provide a brief overview of some of the findings we already obtained. By doing so, we wish to confirm our hypotheses and, at the same time, to prove the feasibility of the methodology we conceived. For the sake of concision, we shall restrict our analyses here to the length of sentences, conjunctions (coordinators and subordinators) and pronouns; but the results for the pronoun "on" and the verbal tenses are very much alike.

[1] The final corpus that we will use for our future analyses is composed of 3,595 requirements (163,000 words) from two different projects.

First, a quantitative analysis is needed to determine whether the requirements corpus really is different from the comparison corpora.

Figure 2 shows the proportion of sentences containing more than twenty-five words in relation to the total number of sentences in each of the three corpora.

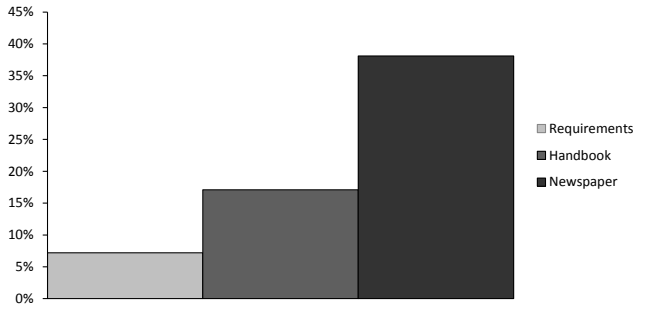

Fig. 2. Proportions of long sentences in the three corpora.

As can clearly be seen, although "long" sentences are present in the requirements (the longest one contains over seventy words), sentences tend to be significantly shorter in requirements (the average is eleven words per sentence) and much longer in newspaper articles (with an average of twenty-four words); the handbook is an intermediate case.

Thanks to a morphosyntactic analysis, we were able to retrieve and count all the occurrences of the conjunctions and the pronouns in the three corpora. For each of them, proportions in relation to the total number of words are shown in Fig. 3.

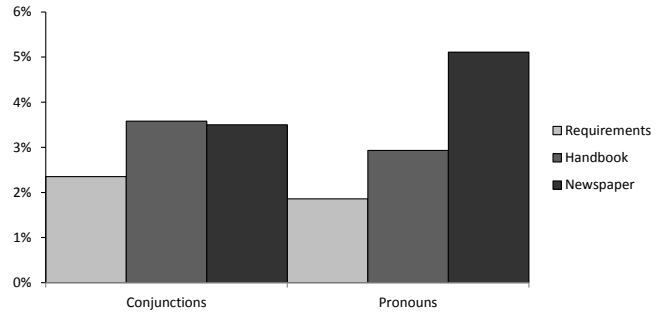

Fig. 3. Proportions of conjunctions and pronouns in the three corpora.

Results are the same as for the length of sentences (with the exception of the handbook corpus containing more conjunctions than the newspaper corpus; in effect, it has more coordinating conjunctions, but less subordinating conjunctions). Therefore, we believe them to be an argument for the existence of a textual genre of requirements, distinct from the genres "technical handbook" and "written press".

Then, we propose a more qualitative analysis by reviewing a selection of examples (all translated from French into English) of requirements containing conjunctions and pronouns found in the reference corpus, and by trying to decide whether or not they should be avoided (as recommended by INCOSE).

Some conjunctions cannot be avoided: "The generator of TCH will check *that* the value of the field PHASE is between 0 *and* FREQ_DIV-1". In this example, the subordinating conjunction "that" is mandatory, since it introduces the dependent clause (this example may be discussed for English, but in French, the complementizer "que" must always be used), and so is the coordinating conjunction "and" to set the lower and higher limits of the interval.

Some conjunctions could be avoided, but they prevent repetition and multiple sentences: "Fields SM_ID *and* FM_ID will be extracted from the BDS". Two sentences would be necessary to avoid the conjunction "and", that would differ by only a single character. As a result, the reader might not notice the difference and assume it is merely a duplicated sentence.

Some conjunctions provide logical information: "for n=2 the size rule is always respected, *but* the 'empty FIFO' test is still required". Thanks to the connector "but", the reader can be certain that the test is required in all cases.

Some conjunctions are not justified: "The requests are to be entered on the FOS *and* the ARPE software manages conflicts between the requests from Spot, Hélios and Pléiades". Here, there is no reason not to make separate sentences.

Following these observations, in a first approximation, we could propose a more precise (and less constraining) rule: for example, the conjunction "and" is to be avoided if it joins independent clauses without a common element (subject, verb or object).

But other problems can also arise because of the absence of the proper coordinator: "This order is rejected *if*: - the automatic NORM mode is active; - the satellite is in MAN mode; the satellite is not in converged mode (GAO or SUP); - a MAN/CAP instruction is already waiting to be executed". In this case, should the order be rejected if one of the following conditions is met ("or"), or only if they are all met ("and")?

Regarding pronouns, many of them are not mandatory, but prevent repetitions of words without being a possible source of ambiguity: "The packet will be generated only if *it* is activated by the LVC". This sentence would not look natural if the noun phrase "the packet" were repeated in full instead of "it".

Some pronouns, however, should be avoided, because otherwise the requirement is no longer autonomous: "*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". This requirement cannot be understood by itself, because the pronoun "it" refers to the subject defined in the previous requirement.

Based on these examples, we could state that a personal pronoun should be used only if it has one and only one possible antecedent in the requirement.

VII. Conclusions

In this research abstract, we tried to emphasize the important issues related to the use of natural language, in particular in the field of requirements engineering.

Controlled Natural Languages are a proposed solution to limit problems such as ambiguity, vagueness or incompleteness: some of them take the form of univocal mathematical notations, but others try to remain close enough to everyday language (even if they are more restrictive on several aspects).

Our goal is to propose a CNL for requirements writing at the French Space Agency (and possibly for other companies). Since no linguistic norm is currently imposed to the engineers, we have decided to consider genuine examples, extracted from the specifications of a space project. The underlying objective is to keep the resulting CNL compatible with the actual practice. Indeed, we have shown that some of the rules imposed by certain standards cannot be literally applied, because they are sometimes too restrictive and sometimes insufficiently so, and that their justifications are not always clear. In this context, we claim that domain experts should collaborate with linguists to specify fine-grained rules.

For this purpose, we have proposed a methodology based on corpus linguistics that combines two approaches: corpus-based and corpus-driven research. This allowed us to formulate a series of hypotheses that can afterwards be tested on our corpora (one corpus of requirements and two comparison corpora).

We proposed a sample of our results that tend to show the existence of a textual genre specific to the reference corpus. We further reviewed some examples to determine to what extent the existing specifications comply with two rules from the recent *Guide for Writing Requirements* issued by INCOSE and also how these rules could be refined.

In our future work, we intend to propose and verify more hypotheses (e.g. concerning the passive voice or the negation). In particular, we will make a more systematic use of text mining tools to identify the regular patterns that compose the grammar of the genre.

VIII. Publications Related to the Thesis

[A] A. Condamines and M. Warnier, "Linguistic Analysis of Requirements of a Space Project and Their Conformity with the Recommendations Proposed by a Controlled Natural Language," in Controlled Natural Language: 4th International Workshop (CNL 2014), B. Davis, K. Kaljurand, and T. Kuhn, Eds. Springer International Publishing, 2014, pp.33-43.

[B] A. Condamines and M. Warnier, "Towards the Creation of a CNL Adapted to Requirements Writing by Combining Writing Recommendations and Spontaneous Regularities: Example in a Space Project," unpublished (submitted to the journal Language Resources and Evaluation in December 2014).

An exhaustive list of my scientific production to date can be found at <http://w3.erss.univ-tlse2.fr/membres/warnier/#publications>.

Acknowledgment

I would like to gratefully acknowledge my Ph.D. advisor, Anne Condamines, as well as Daniel Galarreta and Jean-François Gory (among other people at CNES), for their strong involvement in my thesis, which is fully granted by CNES and the Regional Council of Midi-Pyrénées (France).

I also thank the two anonymous reviewers for their very helpful and motivating evaluations.

References

[1] B. Nuseibeh and S. Easterbrook, "Requirements Engineering: A Roadmap," in *Proceedings of the Conference on The Future of Software Engineering*, New York, NY, USA, 2000, pp. 35–46.

[2] G. J. Pace and M. Rosner, "A Controlled Language for the Specification of Contracts," in *CNL 2009 Workshop*, Marettimo, 2010, pp. 226–245.

[3] E. Kamsties and B. Peach, "Taming ambiguity in natural language requirements," in *Proceedings of the Thirteenth International Conference on Software and Systems Engineering and Applications*, 2000.

[4] Q. Zhang, "Fuzziness - vagueness - generality - ambiguity," *Journal of Pragmatics*, vol. 29, no. 1, pp. 13–31, Jan. 1998.

[5] N. Carlson and P. Laplante, "The NASA automated requirements measurement tool: a reconstruction," *Innovations Syst Softw Eng*, vol. 10, no. 2, pp. 77–91, Sep. 2013.

[6] B. Meyer, "On Formalism in Specifications," *IEEE Softw.*, vol. 2, no. 1, pp. 6–26, Jan. 1985.

[7] T. Kuhn, "A Survey and Classification of Controlled Natural Languages," *Computational Linguistics*, vol. 40, no. 1, pp. 121–170, 2014.

[8] P. Clark, W. R. Murray, P. Harrison, and J. Thompson, "Naturalness vs. Predictability: A Key Debate in Controlled Languages," in *CNL 2009 Workshop*, Marettimo, 2010, pp. 65–81.

[9] AeroSpace and Defence Industries Association of Europe, "Simplified Technical English. Specification ASD-STE100. International specification for the preparation of maintenance documentation in a controlled language. Issue 4." Jan-2007.

[10] International Council on Systems Engineering, "Guide for Writing Requirements." INCOSE, 2011.

[11] V. K. Bhatia, *Analysing genre: Language use in professional settings*. London: Longman, 1993.

[12] H. Somers, "An Attempt to Use Weighted Cusums to Identify Sublanguages," in *NeMLaP3/CoNLL 98 : New Methods in Language Processing and Computational Natural Language Learning*, 1998, pp. 131–139.

[13] L. Guespin, "Normaliser ou standardiser?," *Le Langage et l'homme*, vol. 28, no. 4, pp. 213–222, 1993.

[14] F. Gaudin, *Pour une socioterminologie : des problèmes sémantiques aux pratiques institutionnelles*. 1993.

[15] D. Hymes, "Models of the Interaction of Language and Social Setting," *Journal of Social Issues*, vol. 23, no. 2, pp. 8–28, Apr. 1967.

[16] E. Tognini-Bonelli, *Corpus Linguistics at Work*. John Benjamins Publishing, 2001.

[17] D. Biber, "Corpus-Based and Corpus-driven Analyses of Language Variation and Use," in *The Oxford Handbook of Linguistic Analysis*, 1st ed., B. Heine and H. Narrog, Eds. Oxford University Press, 2009.

[18] S. O'Brien, "Controlling Controlled English. An Analysis of Several Controlled Language Rule Sets," in *Proceedings of EAMT-CLAW*, 2003, pp. 105–114.

[19] S. Bouquet, "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*, no. 56, pp. 143–156, Jun. 2007.

[20] D. Malrieu, "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*, no. 56, pp. 157–178, Jun. 2007.