Tailoring a Controlled Language Out of a Corpus of Maintenance Reports

Yannis Haralambous
UMR 6285 Lab-STICC, DECIDE
& Département Informatique,
IMT Atlantique,
CS 83818,
29238 Brest Cedex 3
yannis.haralambous@imt-atlantique.fr

Tian Tian
UMR 6285 Lab-STICC, DECIDE
& Département Informatique,
IMT Atlantique,
CS 83818,
29238 Brest Cedex 3
tian.tian@imt-atlantique.fr

Abstract

We introduce a method for tailoring a controlled language out of a specialized language corpus, as well as for training the user to ensure a smooth transition between the specialized and the controlled language. Our method is based on the selection of maximal coverage syntax rules. The number of rules chosen is a naturalness vs. formality parameter of the controlled language. We introduce a training tool that displays segmentation into left-to-right maximal parsed sentences and allows utterance modification by the user until a complete parse is achieved. We have applied our method to a French corpus of maintenance reports of boilers in a thermal power station and provide coverage and segmentation results.

1 Introduction

The distinction between naturalist and formalist approach to controlled natural language has been widely discussed in the literature (Pool, 2006; Clark et al., 2010; Gruzitis et al., 2012; Marrafa et al., 2012; Kuhn, 2014). We will adopt an intermediate approach. Indeed, in this paper we deal with the specific problem of optimizing information and knowledge extraction out of utterances in a specialized but spontaneously written language, the language of maintenance reports of the boilers of a thermal power station. The reports are written under conditions of stress and lack of time, and therefore largely adopt a "telegraphic" spontaneous style, without post-editing or spell checking. The ultimate goal of our ongoing project is to mine the text in order to have it correlated with time-stamped data from sensors in the equipment, in search of anomalies. Use of a controlled language is expected to improve the text mining process, but asking technicians to rigorously adhere to a specific controlled language fragment is not an option.

We have therefore chosen to make a compromise by designing a controlled language based on the existing maintenance reports corpus vocabulary but with a restricted grammar. We have developed an editor that displays, in a non-intrusive way, success or failure of the syntax parse of written utterances. This means that we are adopting a naturalist approach while using tools from the formalist approach (subsets of syntax rules) to simplify the legacy natural language and make it easier to interpret.

To bring the controlled language closer to standard French, we used an additional, totally independent corpus, consisting of sentences in regular and carefully edited French. We extracted Phrase-Structure Grammar rules from this corpus.

Let us call $\mathcal{S}$ the maintenance report corpus and $\mathfrak{S}$ its set of production rules, $\mathcal{M}$ the regular French corpus and $\mathfrak{M}$ its production rules. By using the vocabulary of $\mathcal{S}$ and allowing only the most frequent elements of $\mathfrak{M}$ (and potentially some frequent rules from $\mathfrak{S}$) we define a controlled language that is a simplification of $\mathcal{S}$ (Saggion, 2017). Our first innovation is the possibility of tailoring the formality/naturalness of the controlled language by reducing/increasing the number of allowed production rules. Furthermore, the fact that these rules belong to a "golden corpus of standard French," entails a regularization of the informal (and syntactically chaotic) language of $\mathcal{S}$.

Syntax is very central to our approach because the vocabulary and its morphological variation are limited, due to the technical nature of the corpus.

Our second innovation is an editor with a "non-intrusive" training interface. A parsed utterance is displayed in blue color (or bold style, or some other graphical attribute) and the potentially unparsed part of it remains in standard style. Depending on working conditions during the text authoring act, the technician can choose to invest time and energy in "improving" eir[1] linguistic production, or ignore

[1] We use gender-neutral Spivak pronouns https://en.wikipedia.org/wiki/Spivak_pronoun.

the fact that the utterance has not been entirely parsed. When modifying the utterance, immediate feedback (by some graphical artefact) is provided to the author, who is thereby entering a smooth training process.

2 Related work

(Clark et al., 2010) define CPL and CPL-Lite as two variants of the same *Computer-Processable Language*, where

> While CPL searches for the best interpretation, CPL-Lite is simpler and interpreted deterministically (no search or use of heuristics). (Clark et al., 2010, 69)

In CPL-Lite, 113 sentence templates are allowed, giving rise to an equal number of binary predicates in Prolog-like syntax. We generalize this principle by allowing a variable number of production rules.

Despite the differences between the two languages, (Kuhn, 2014) considers only one CPL language and assigns it a PENS classification of $P^3E^3N^4S^2$FWI.[2] Other controlled languages based on a limited number of production rules are SQUALL ($\sim$ 50 rules) (Ferré, 2012), ucsCNL ($\sim$ 140 rules) (Barros et al., 2011) and Attempto ($\sim$ 360 rules without disjunctions) (Fuchs, 2018). These have been classified as $P^5E^2N^3S^4$FWA, $P^5E^2N^4S^4$FWDA and $P^4E^3N^4S^3$FWA, respectively by (Kuhn, 2014).

As we see, the lesser the rules (e.g., in SQUALL) the higher is P (precision). In our case the controlled language can be built with a variable number of production rules, so P can be variable, probably between P^2 and P^4. Expressiveness is rather low for the languages mentioned, but in our case this is of no concern since the application domain is very narrow: quantification is very sparse since maintenance reports concentrate on a small number of boilers and their parts, general rule structures are also limited since sentences are almost always declarative—only the presence of negation is mandatory, to express failure of equipment. As for naturalness, by using the same vocabulary as $\mathcal{S}$ and building syntax based on the rules of standard well-formed French, a high degree of naturalness is achieved, which we estimate around N^4. Simplicity can be assessed with more difficulty since there is no explicit description of the language. This description would imply giving and explaining all production and semantic rules involved and such a description can indeed be done but will not be provided to the language's users. Users are intended to adapt progressively to the controlled language—potentially a short notice on the editor's working principle may be addressed to them, but it will by no means be a comprehensive description of the language. We would therefore rather consider this language as S^2, a "language without exhaustive description," even though such a description would theoretically be possible. As for properties, these would be W (written) D (specific narrow domain) and I (industry).

[2] (Kuhn, 2014) defines letter codes for properties of controlled natural languages of different categories. In this frame, C stands for comprehensibility as goal of the language; T for translation; F for formal representation. As for the form of the language, W stands for written languages and S for spoken ones; and D stands for languages in a specific narrow domain. As for origin of the controlled language, three codes are defined: A stands for languages originating from academia, I from industry and G from government. (Kuhn, 2014) defines the PENS classification scheme to describe controlled languages according to four axes: P (precision), E (expressiveness), N (naturalness) and S (simplicity). For each dimension, five degrees (arbitrarily) are used, such as $P^1E^5N^5S^1$ for standard English and $P^5E^1N^1S^5$ for propositional logic.

3 Description of the Corpora

Our main corpus $\mathcal{S}$ consists of 2,280 maintenance reports, written in 8-hour intervals during two years. The volumetry of $\mathcal{S}$ is as follows: 30,851 sentences, 138,140 words. We explore its properties in Sections 4 (lexicon), 5 (morphology) and 6 (syntax).

To serve as a "ground truth" of French syntax, we built a second corpus, $\mathcal{M}$, based on eleven Harlequin-like novels by a well-known author. They are written in informal everyday French language, carefully edited by the publisher since the given novels are best-sellers with a very large audience. We have parsed the two corpora using the Stanford CoreNLP parser and have kept only syntax trees. On the syntax level, $\mathcal{M}$ provides mostly short to medium-length sentences with basic syntax. They include informal sentences (in the form of dialogs) but also simply-written formal sentences, so that frequent production rules from $\mathfrak{M}$ can establish a transition from informal to relatively formal utterances in the maintenance reports. Using a legacy corpus such as FTB (Abeillé et al., 2003) (originating from *Le Monde* articles) instead of $\mathcal{M}$ would be inadequate in our case, because of FTB's high syntactic complexity that is unlike the average syntax of $\mathcal{S}$ sentences.

4 The Lexical Level

In order to parse $\mathcal{S}$ efficiently we have preprocessed the text and extracted codes, abbreviations and equipment identifiers. We replaced these forms during parsing by a unique mark to avoid misinterpretations. We also detected misspelled/alternatively spelled words and replaced them by standard forms. As for $\mathcal{M}$, we removed sentences with an elliptic syntax (titles, sentences ending with ellipsis, etc.) using heuristic filters. After filtering we kept 48,693 sentences (650,847 words).

To evaluate $\mathcal{S}$'s vocabulary we have randomly chosen a subset $\mathcal{M}' \subset \mathcal{M}$, having the same volumetry as $\mathcal{S}$. Unsurprisingly, $\mathcal{S}$ has a significantly more restricted vocabulary than $\mathcal{M}'$: 5,505 different lexemes in the former vs. 9,374 in the latter. Their distribution is as follows:

	ADJ	NOUN	VERB	VN	PROPN
$\mathcal{S}$	7,137	**43,034**	12,616	9,295	**18,305**
$\mathcal{M}'$	**9,224**	36,610	**26,335**	**23,737**	8,358

where VN denotes past participles and PROPN proper nouns to which we added codes, abbreviations and equipment identifiers. We see that $\mathcal{S}$ has clearly more "proper nouns" and slightly more nouns than $\mathcal{M}'$, but all other parts of speech are underrepresented.

When words in $\mathcal{S}$ happen to be both frequent and complex, they occasionally undergo significant variation. Let us take the example of word "régénération," the sixth most frequent noun in $\mathcal{S}$ (485 occurrences in its standard form), which appears in the following alternative forms:

> régé: 1,117 times (apocope)
> Régé: 549 times (apocope)
> Rége: 14 times (apocope & accent error)
> rége: 12 times (apocope & accent error)
> rege: 11 times (unaccented apocope)
> Regé: 6 times (apocope & accent error)
> Rege: 5 times (unaccented apocope)
> régés: 5 times (apocoped plural)
> regé: 4 times (apocope & accent error)
> Regénération: twice (accent error)
> Régénaration, Régénèration, regenaration, regeneration, régeneration, régéneration: hapaxes (accent or spelling errors).

Variation is also frequent in English-origin words such as "bypass":

> bypass: 240 times
> by-pass: 147 times (with hyphen)
> Bypass: 19 times (capitalized)
> ByPass: twice (camel notation)
> By-pass, By-Pass, BY-pass: hapaxes.

Some abbreviation processes are peculiar such as the contraction "ppe" (for word "pompe") that occurs 117 times in the singular and 11 times in the plural number, or the apocope "échaff" (7 times) based on an erroneous ($\rightarrow$ two 'f's) spelling of word "échafaudage".

We encountered 920 cases of erroneous/non-standard spellings, involving 3,772 occurrences (out of which 588 were hapaxes).

The vocabulary is technical and has to remain unreduced. However, an interactive spell-checking and auto-completion device can be useful to avoid ambiguities, like in the cases of apocopes "aéro" or "régul" that can have a multitude of completions.

5 The Morphological Level

French is an uncased language, so that its morphological variation is focused mainly on conjugation for verbs and (less importantly) on number and gender of nouns and adjectives.

The use of verbs is very restricted in $\mathcal{S}$. While in $\mathcal{M}'$ we encountered 24 frequent different combinations of mode, tense, number and person, not counting infinitives and participles, in $\mathcal{S}$ there were only three frequent ones:

	P3s	P3p	F3s
$\mathcal{S}$	2,638	56	**106**
$\mathcal{M}'$	**3,524**	**159**	97

where P stands for present and F for future tense, 3 for third person and s/p for singular/plural. Episodic detection of other verb forms is often due to misspellings, such as in

> appel astreinte GN pour information que l'astreinte électrique ne peux rien faire de plus aujourd'hui !!!

where the P1s form of verb "pouvoir" is mistakenly used instead of the P3s form.

The low morphological variation of the $\mathcal{S}$ corpus comes as no surprise since maintenance reports use P3s and P3p to communicate the state of one or more devices at the time of report writing, and F3s (or F3p) for interventions that are scheduled in the near future, as in:

> la fin de la régénération de la chaîne 2 se terminera vers 7h45

6 The Syntax Level

Because of the conditions under which maintenance reports are authored, we notice a predominance of the "telegraphic style". This results in two phenomena: (1) elliptical language, as many obvious words are omitted for the sake of brevity; (2) chaotic syntax, where elementary rules of French sentence construction are broken. Typical examples are:

(1) fuite impulsion séparateur stable

which is a sentence containing neither verb nor determinant, consisting of three nouns and an adjective, and

(2) Faire avis sur fuite d'huile sulzer que si en augmentation voir consigne MPy

which seems like the (unpunctuated) transcription of an oral utterance. Here are completed version of these utterances, including implicit intentions, missing determinants and verbs:

(1′) *Nous avons constaté que la* fuite *de l'*impulsion *du* séparateur *est* stable.

(2′) *Il est conseillé de* faire *un* avis sur *la* fuite d'huile *du* sulzer *disant que si elle est en* augmentation *alors il faudra* voir *la* consigne *du* MPy.

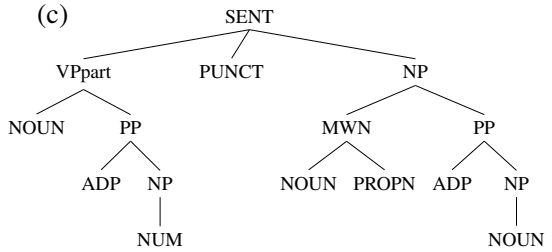
6.1 Parsing

We pre-processed the $\mathcal{S}$ corpus with regular expressions to replace all numerals and physical values with a single NUM tag and all device names and abbreviations with the PROPN tag: this has cut the number of distinct sentences in half, going from 30,851 sentences in the original corpus to 15,065. We then parsed the data using Stanford CoreNLP parser in order to obtain Phrase-Structure Grammar syntax trees of both corpora. We removed lexical leaves. We then extracted all production rules. Here is an example of this process:

(a) Arrêt à 20h00, arrêt TG1 à 20h15

(b) Arrêt à NUM , arrêt PROPN à NUM

(c)



(d) SENT → VPpart PUNCT NP, VPpart → NOUN PP, PP → ADP NP, NP → NUM, NP → MWN PP, MWN → NOUN PROPN, PP → ADP NP, NP → NOUN.

We calculated the occurrence frequency of every production rule in the corpus: it is the number of sentences in the syntax trees in which it is used (multiple use of a rule in the same syntax tree is not taken into account). The frequency of production rules follows a Zipf distribution: $\mathfrak{S}$ consists of 8,583 rules, the most frequent of which (namely PP → ADP NP) has a frequency of 18,584 and the distribution has a tail of 5,662 hapaxes (66% of the rules). On the other hand, $\mathfrak{M}$ consists of 30,930 rules, the most frequent (once again PP → ADP NP) having a frequency of 48,573 and the tail contains 23,203 hapaxes (75% of the rules). To give an example of the difference between $\mathfrak{S}$ and $\mathfrak{M}$, the SENT → PP rule which is typical of the telegraphic style and corresponds to expressions such as "en service" (frequency 263 in $\mathfrak{S}$) does not appear at all in $\mathfrak{M}$—on the other hand, the NP → DET NOUN rule that corresponds to the fundamental property of French nouns of being preceded by a determinant (a property that is often relaxed in telegraphic style) is the *second* most frequent rule in $\mathfrak{M}$ but only the *eleventh* in $\mathfrak{S}$.

6.2 Frequency-Based Subgrammars

Let T, N be fixed sets of terminals and non-terminals, and S an initial symbol. If R is a set of production rules we denote by $G(R)$ the corresponding formal grammar and by $L(R)$ the formal language recognized by $G(R)$. When $R \subset R'$ (while T, N, S remain fixed) then, obviously, $L(R) \subset L(R')$. Therefore by allowing a subset of production rules we obtain a sub-language. By keeping only the most frequent rules, we allow only for the most common syntactic features in the sub-language, and thereby the sub-language becomes a *simplified version* of the original language (Saggion, 2017, Ch. 4). Keeping a strongly reduced set of rules allows efficient manual definition of semantic rules, according to the principle of compositionality (Partee, 1995; Bird et al., 2019). The more production rules we allow, the more cumbersome it is to define the corresponding semantic rules. It is impossible to define semantic rules for an entire natural language, but it is possible to do so for controlled languages, provided their set of production rules is of reasonable size.

Input : Sentence $(w_1, \ldots, w_n)$, rules R
Output : Segments $(s_1, \ldots, s_m)$ where
$s_i := (w_{\ell(i)}, \ldots, w_{r(i)})$ for
$1 \leq i \leq n$, and
rest $:= (w_{n-j}, \ldots, w_n)$ for $j \geq 0$,
or $\emptyset$

if $(w_1, \ldots, w_n) \in L(R)$ **then**
  $s_1 := (w_1, \ldots, w_n)$;
  return $((s_1), \emptyset)$
end
else if $\nexists i, 1 \leq i < n$ *such that* $(w_1, \ldots, w_i) \in L(R)$ **then**
  return $(\emptyset, (w_1, \ldots, w_n))$
end
else
  $k \leftarrow 1$;
  $\ell(k) \leftarrow 1$;
  while $\exists r(k), \ell(k) \leq r(k) \leq n$ *such that* $(w_{\ell(k)}, \ldots, w_{r(k)}) \in L(R)$ **do**
    $s_k \leftarrow (w_{\ell(k)}, \ldots, w_{r(k)})$;
    $\ell(k+1) \leftarrow r(k) + 1$;
    $k \leftarrow k + 1$;
  end
  $k \leftarrow k - 1$;
  if $r(k) = n$ **then**
    return $((s_1, \ldots, s_k), \emptyset)$
  end
  else
    return $((s_1, \ldots, s_k), (w_{r(k)+1}, \ldots, w_n))$
  end
end

Algorithm 1: Left-Right Maximal Segmentation Algorithm

In our case, production rules are ordered by decreasing frequency and we can consider sets such as $\mathfrak{M}_{\geq 50}$: "the language produced by the terminals and non-terminals of the $\mathcal{M}$ corpus as well as the set of rules of frequency greater or equal to 50"; or $\mathfrak{S}_{\geq 2}$: "the language produced by the terminals and non-terminals of the $\mathcal{S}$ corpus using rules that are not hapaxe"s; etc.

We will use these sets as a base for tailoring controlled languages with a variable trade-off between formality and naturalness.

6.3 Left-Right Maximal Segmentation

According to (Angelov and Měchura, 2018), editors for controlled languages are

> of roughly two types. The first is the so called syntax editors which let the user manipulate a logical structure, while the actual text is just a byproduct. [...] The second kind is called predictive editors, which opt to work directly on the text level and guide the user by showing the set of possible continuations.

In the context of our project, both editor types are doomed to fail: boiler maintenance technicians under strong stress are probably not keen on visualizing syntax trees of their utterances, and a predictive editor would be incompatible with the high speed (not to say, haste) of the authoring act. Indeed, a technician having important information about the status of the equipment to transmit should be able to do so without any interference, and if there is time for improvement this can only happen a posteriori, after the authoring act is completed.

So the question is: how can we train technicians into using the controlled language in a way that is acceptable under the circumstances? The least intruding way would be to have a simple color code indicating successful/unsuccessful parsing, but then we fall into the other extreme: no information is given to the user on how to improve eir utterances, which can result in frustration when possible corrections have to be guessed.

The intermediate solution we adopt is to display a segmentation of the utterance into parsed sentences and, potentially, an unparsed rest. The rationale of this solution (loosely based on the pumping lemma for context-free languages) is that if an utterance (and in particular, a long one) is not a sentence for the controlled language, then there is a high probability (see Table 1) that some contiguous subsegments of it are nevertheless recognized as sentences. Starting from the left we mark the largest part of the utterance that is a sentence and iteratively repeat this process for the rest of the utterance, until we have reached a maximum sequence of segments that are sentences for the controlled language (see Alg. 1).

This gives the technician an understanding on how the utterance is decomposed into sentences by the parser. If the entire utterance is recognized by the parser, the author can leave it as such, otherwise e can intervene to change the phrase structure and attempt validation anew. If some words remain unparsed after the last segment, the author can attempt to incorporate them into the last segment, or to add text to produce a complete sentence out of them.

The success of this "training process" will depend on the coverage of the grammar. In Table 1

Table 1: Results of Maximal Left-Right Segmentation of the $\mathcal{S}$ Corpus

$\mathfrak{M}$ \ $\mathfrak{S}$		≥ 2 (1,411 rules)	≥ 3 (762 r.)	≥ 5 (412 r.)	≥ 10 (163 r.)	≥ 50 (21 r.)	$\emptyset$ (0 r.)
		Coverage (sentences with at least one segment)					
≥ 5	(2,204 rules)	90.87%	80.86%	78.8%	74.95%	64.93%	37.94%
≥ 10	(1,391 rules)	88.68%	77.11%	74.67%	70.47%	58.29%	28.72%
≥ 50	(460 rules)	81.72%	67.71%	64.84%	59.95%	43.38%	13.89%
≥ 100	(272 rules)	75.88%	61.18%	58.43%	53.48%	34.48%	6.96%
≥ 500	(81 rules)	61.36%	44.06%	42.08%	36.8%	18.15%	5.14%
		Rest (ratio between ratio length and utterance length)					
≥ 5	(2,204 rules)	0.07	0.08	0.09	0.11	0.16	0.25
≥ 10	(1,391 rules)	0.1	0.12	0.14	0.16	0.2	0.31
≥ 50	(460 rules)	0.19	0.23	0.25	0.28	0.33	0.46
≥ 100	(272 rules)	0.25	0.29	0.32	0.35	0.38	0.48
≥ 500	(81 rules)	0.42	0.46	0.48	0.5	0.48	0.51
		Average segment size (in words)					
≥ 5	(2,204 rules)	5.34	5.97	6.1	6.3	6.52	7.51
≥ 10	(1,391 rules)	5.63	6.18	6.31	6.51	6.73	7.88
≥ 50	(460 rules)	6.19	6.56	6.74	6.97	7.13	8.81
≥ 100	(272 rules)	6.39	6.7	6.88	7.09	7.22	10.2
≥ 500	(81 rules)	7.04	7.24	7.32	7.41	7.18	10.7

we present results of maximal left-right segmentation to the $\mathcal{S}$ corpus. Lines represent the various sets of $\mathfrak{M}$ rules used to segment the corpus. Taking all non-hapax rules we obtain the best results, but we must deal with semantic rules for thousands of syntax rules—on the other hand, when using only rules of high frequency, coverage drops drastically but so does the number of rules. The foremost right column represents the case where only $\mathfrak{M}$ rules are used to define the controlled language, the other columns consider the case where some $\mathfrak{S}$ rules are also allowed. The worst result occurs on the 5th line when only 81 $\mathfrak{M}$ rules and no $\mathfrak{S}$ rules are used: these conditions result in a very strict controlled language and it is not surprising that only 5.14% of the sentences of the existing corpus provide a segment. We can call $\mathfrak{M}_{\geq 500}$ the "strict strategy," where one wishes a syntactically simple controlled language at all cost.

Another strategy is $\mathfrak{M}_{\geq 5}$ which represents a significant effort to keep the controlled language close to standard French language. It involves preparing semantic rules for 2,204 syntactic rules, which is a considerable task. Using this approach, if technicians keep on writing as they did in the $\mathcal{S}$ corpus, in 37.94% of cases they will get a segmentation of, in average, three fourths of the utterance, with segments of an average length of 7.51 words. This is the "$\mathfrak{M}$-only at all costs" strategy.

A third strategy is to allow for additional rules coming from $\mathfrak{S}$ (note that the number of $\mathfrak{S}$ rules in the table stands for rules not already included in $\mathfrak{M}$). By taking a small number of $\mathfrak{S}$ rules, for example 21 rules of frequency ≥ 50, coverage increases significantly: we reach 64.93% vs. 37.94% in the previous strategy.

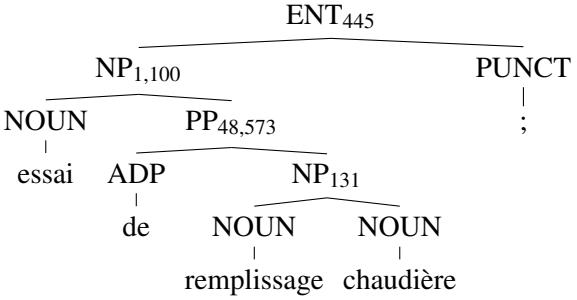
Finally the "most expensive" extreme is to take ≥ 2 rules from $\mathfrak{S}$ and ≥ 5 rules from $\mathfrak{M}$, which makes a total of 2,615 rules to manage, with a coverage of 90.87% of sentences and segments covering 93% of each sentence, in average. We consider this approach as a kind of overfitting, where one aims to reproduce the chaotic nature of the legacy language in a controlled environment, at a very high cost. Fortunately many intermediate solutions exist between these extremes.

In the following we will give examples of various utterances belonging or not to the controlled language for specific parameters.

7 Examples

7.1 $\mathfrak{M}_{\geq 50}$ and no $\mathfrak{S}$ rules

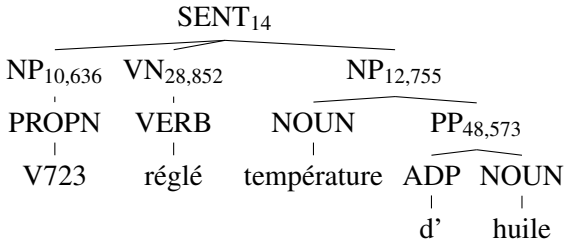
Our first example[3] is the one of a sentence that is successfully parsed with rule in $\mathfrak{M}_{\geq 50}$:



As we see the rule with lowest frequency is NP → NOUN NOUN, which can be found in $\mathcal{M}$ in expressions such as [[samedi]$_{\text{NOUN}}$ [après-midi]$_{\text{NOUN}}$]$_{\text{NP}}$.

7.2 $\mathfrak{M}_{\geq 10}$ and no $\mathfrak{S}$ rules

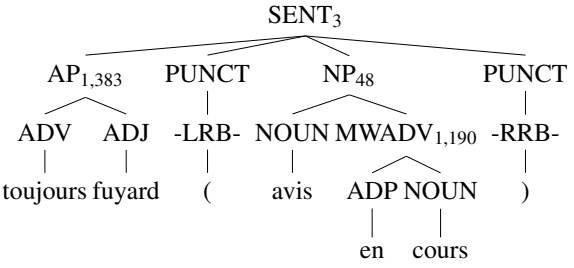
The following example uses $\mathfrak{M}$ rules of frequency ≥ 10 (with at least one rule of frequency ≤ 50) :



The least frequent rule here is SENT → NP VN NP, which appears only 14 times in the $\mathfrak{M}$ corpus, in syntagms such as [[Le coup de poing]$_{\text{NP}}$ [partit]$_{\text{VN}}$ [tout seul]$_{\text{NP}}$]$_{\text{S}}$

7.3 $\mathfrak{M}_{\geq 2}$ with no $\mathfrak{S}$ rules

The following example stretches syntax at its limits since it uses rare $\mathfrak{M}$ rules (of frequency higher than 2 but less than 10):

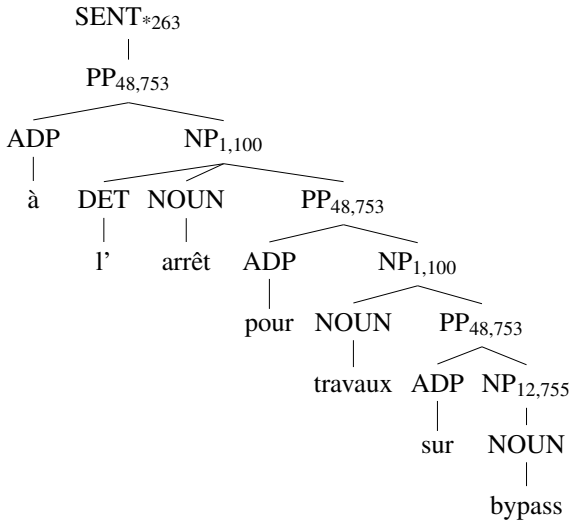


SENT → AP PUNCT NP PUNCT is a rare rule that appears almost by accident in a sentence such as [[Salut]$_{\text{AP}}$ [,]$_{\text{PUNCT}}$ [Marko]$_{\text{NP}}$ [?]$_{\text{PUNCT}}$]$_{\text{S}}$, which is hardly similar to our example. This sentence is clearly of telegraphic style. The rule NP → NOUN MWADV is not frequent either, it is attested in syntagms such as [[oui]$_{\text{NOUN}}$ [bien sûr]$_{\text{MWADV}}$]$_{\text{NP}}$.

[3] Indices in the syntax trees denote frequency in $\mathfrak{M}$, and starred indices denote frequency in $\mathfrak{S}$.

7.4 $\mathfrak{M}_{\geq 50} \cup \mathfrak{S}_{\geq 10}$

We now turn to an example that cannot be parsed entirely with $\mathfrak{M}$ rules and requires at least one $\mathfrak{S}$ rule, of frequency higher than 10:



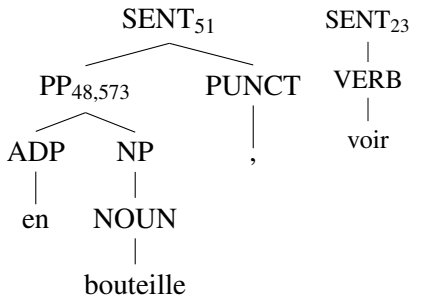
The rule from the $\mathcal{S}$ corpus is SENT → PP ("a sentence can be a prepositional phrase," which is typically telegraphic style) and it appears 263 times in $\mathcal{S}$. All other rules are quite frequent in $\mathfrak{M}$.

8 The Editor

To train users of the controlled language we have developed a device that parses word sequences on-the-fly, and displays maximal parsed segments using blue color (or some other graphical style) and brackets, from left to right. For example, for the utterance "en bouteille, voir schéma," the user will progressively see the following:

en
[en bouteille]
[en bouteille,]
[en bouteille,] [voir]
[en bouteille,] [voir] schéma

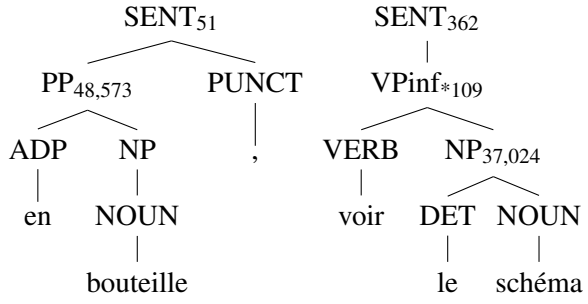
The last word cannot be absorbed by a segment if, e.g., only $\mathfrak{M}_{\geq 20} \cup \mathfrak{S}_{\geq 10}$ rules are allowed. This segmentation is based on the following two trees:



The second tree uses a rare rule SENT → VERB, which is of frequency 23 in $\mathfrak{M}$. At this point, the user can attempt a correction by adding a definite article "le" between "voir" and "schéma":

[en bouteille,] [voir le schéma]

The color changes to blue as the utterance is now entirely parsed, using the following trees:



The second tree uses the rule VPinf → VERB NP that belongs to $\mathfrak{S}$ with a frequency of 109.

We are planning to test the device and establish performance evaluation through user feedback.

9 Conclusion and Perspectives

We have presented a methodology for tailoring a controlled language out of the lexicon and morphology of a corpus, using the most frequent Phrase-Structure Grammar syntax rules of another corpus. We have applied this approach to a corpus of industrial equipment maintenance reports written in telegraphic style, as the former, and a corpus of Harlequin-like French novels as the latter. The goal is to make user utterances more easily interpretable. By allowing also some syntax rules from the original corpus, we obtain a better result in terms of coverage of utterances by syntax rules. By varying the number of allowed rules from the Harlequin-like novels and from the maintenance reports, we can change the formality/naturalness properties of the controlled language.

We have also presented a new editor type that is neither syntax-displaying nor predictive, but provides the user with information on the best possible left-to-right segmentation of the utterance into parsed sentences. This allows optional intervention by the user in order for the complete utterance to get parsed. The editor is purposely non-intrusive since the conditions under which maintenance reports are written do not always allow for a calm and reasoned reflection on syntax.

This is an ongoing project, the final goal of which is to achieve anomaly detection in reports, eventually correlating (timestamped) textual data with temporal series of data originating from sensors in the boilers, in search of anomalies. For this, formal interpretation of the reports can be useful but is not indispensable since text mining methods can compensate the lack of full interpretation. Another potential application is to correlate linguistic parameters of the corpus with author identities, since these are always provided in the reports. This would allow to evaluate the variability in lexicon, morphology and syntax due to author change.

10 Acknowledgments

This work has been realized in the frame of the European Regional Development Fund project AAP FEDER – LEARN IA, funded by the *Conseil régional de Bretagne* and the *European Union*.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In *Treebanks*, pages 165–187. Kluwer.

Krasimir Angelov and Michal Boleslav Měchura. 2018. Editing with search and exploration for controlled languages. In *Controlled Natural Language*, volume 304 of *Frontiers in Artificial Intelligence and Applications*, pages 1–10. IOS Press.

Flávia A. Barros, Neves Laís, Érica Hori, and Dante Torres. 2011. The ucsCNL: A controlled natural language for use case specifications. In *Proceedings of SEKE'2011, Miami Beach, Florida*, pages 250–253.

Steven Bird, Ewan Klein, and Edward Loper. 2019. *Natural Language Processing with Python*, 2nd edition. https://www.nltk.org/book/.

Peter Clark, William R. Murray, Phil Harrison, and John Thompson. 2010. Naturalness vs. Predictability: A key debate in controlled languages. In *CNL 2009 Workshop*, volume 5972 of *Springer LNAI*, pages 65–81.

Sébastien Ferré. 2012. SQUALL: A controlled natural language for querying and updating RDF graphs. In *CNL 2012*, volume 7427 of *Springer LNCS*, pages 11–25.

Norbert E. Fuchs. 2018. Understanding texts in Attempto controlled English. In *Proceedings of the 6th International Workshop on Controlled Natural Language (CNL 2018)*, volume 304 of *Frontiers in Artificial Intelligence and Applications*, pages 75–84. IOS Press.

Normunds Gruzitis, Peteris Paikens, and Guntis Barzdins. 2012. FrameNet resource grammar library for GF. In *CNL 2012*, volume 7427 of *Springer LNCS*, pages 121–137.

Tobias Kuhn. 2014. A survey and classification of controlled natural languages. *Computational Linguistics*, 40:121–170.

Palmira Marrafa, Raquel Amaro, Nuno Freire, and Sara Mendes. 2012. Portuguese controlled language: Coping with ambiguity. In *CNL 2012*, volume 7427 of *Springer LNCS*, pages 152–166.

Barbara Partee. 1995. Lexical semantics and compositionality. In *An Invitation to Cognitive Science: Language*, volume 1, pages 311–360. MIT Press.

Jonathan Pool. 2006. Can controlled languages scale to the Web? In *CLAW 2006, AMTA 2006: 5th International Workshop on Controlled Language Applications*, pages 1–12.

Horacio Saggion. 2017. *Automatic Text Simplification*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.