Tailoring a Controlled Language Out of a Corpus of Maintenance Reports

Tian TIAN

April 26th 2022

Outlines

1. About Me
2. About Natural Language Processing (NLP)
3. Project Learn Artificial Intelligence (Learn-AI)
4. The Maintenance Corpus
5. Tailoring a controlled language out of the corpus
6. The Editor to help users to get used to the controlled natural language
7. Evaluation
8. Examples from Tailored Controlled Languages
9. Conclusion and Perspectives

Outline

1. About Me
2. About Natural Language Processing (NLP)
3. Project Learn Artificial Intelligence (Learn-AI)
4. The Maintenance Corpus
5. Tailoring a controlled language out of the corpus
6. The Editor to help users to get used to the controlled natural language
7. Evaluation
8. Examples from Tailored Controlled Languages
9. Conclusion and Perspectives

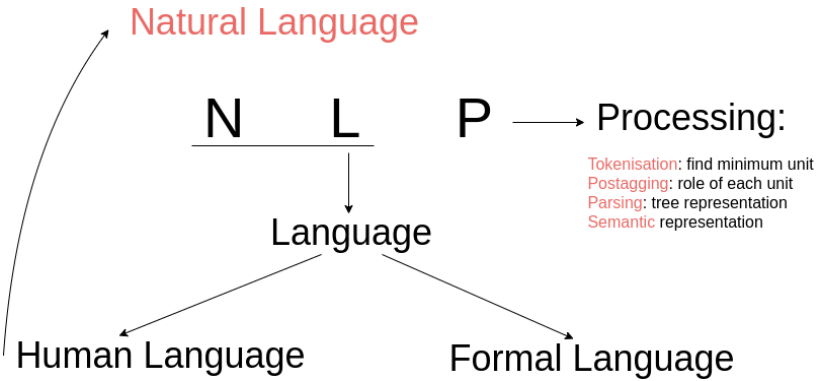
Tian Tian (田甜) Post-doc at IMT Atlantique, Lab-STICC

- PhD degree on Natural Language Processing (NLP) in 2019 at Sorbonne Nouvelle University, Paris, France
  - Thesis CIFRE, E-reputation (Sentiment analysis)
  - Adaptation of well-formed model on raw text
  - Named Entity Recognition, POS-tagging
  - Lexical Normalisation of sociaux media texts
- ATER at Sorbonne Université (2019-2020), Paris, France
- Post-doc at IMT Atlantique since 2021, Brest, Brittany, France
  - Symbolic approach for NLP
  - Heterogeneous data fusion
  - Knowledge extraction

Outline

1. About Me
2. About Natural Language Processing (NLP)
3. Project Learn Artificial Intelligence (Learn-AI)
4. The Maintenance Corpus
5. Tailoring a controlled language out of the corpus
6. The Editor to help users to get used to the controlled natural language
7. Evaluation
8. Examples from Tailored Controlled Languages
9. Conclusion and Perspectives

Explained NLP



NLP is sooooo difficult!

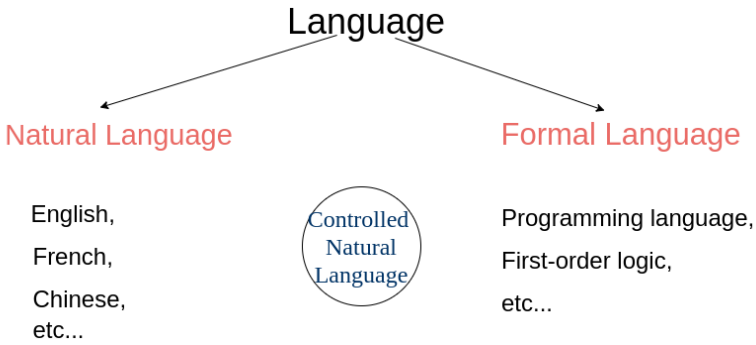
- Tokenisation: ”aujourd’hui” vs ”l’information”
- Meanings of words: ”avocat” (a profession or a vegetable)
- Part-of-speech tagging:

”La	belle	ferme	le	voile.”
DET	NC	VERB	DET	NC
DET	ADJ	NC	PRON	VERB

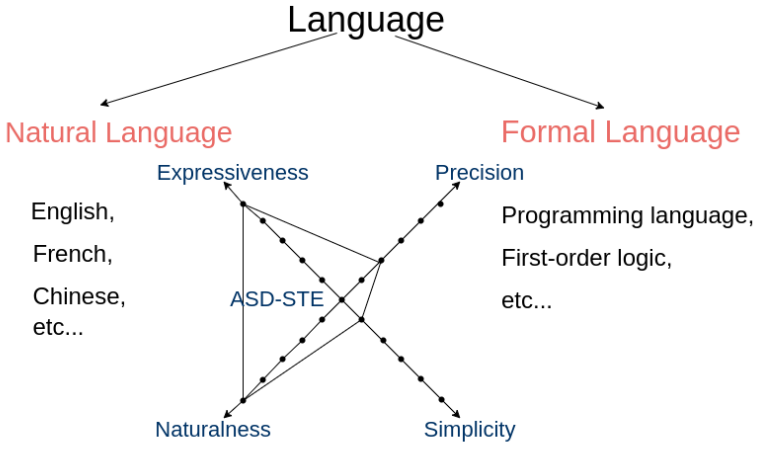
- Parsing: ”I saw the man in the park with a telescope.”
- Pragmatic:

- Do you have the password?
- Yes, I do.
- ...@*%=¡&... So... Can you give it to me?
- Yes, I can.
- ...@*%=¡&... But... WHAT IS IT?
- The password is ...

Between Natural Language and Formal Language

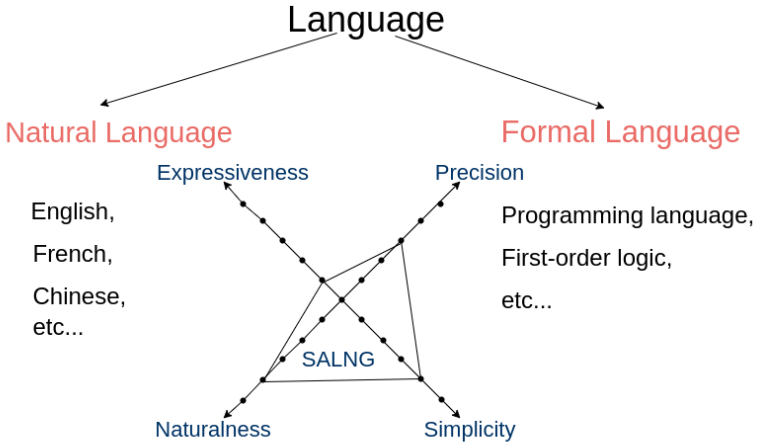


Example of existing Controlled Natural Language: ASD ($P^2E^5N^5S^1$)



ASD-STE: ASD Simplified Technical English (ASD 2013)

Example of existing Controlled Natural Language: SLANG ($P^3E^1N^4S^2$)



SLANG: Standard Language, developped by Ford Motor Company since 1990

Outline

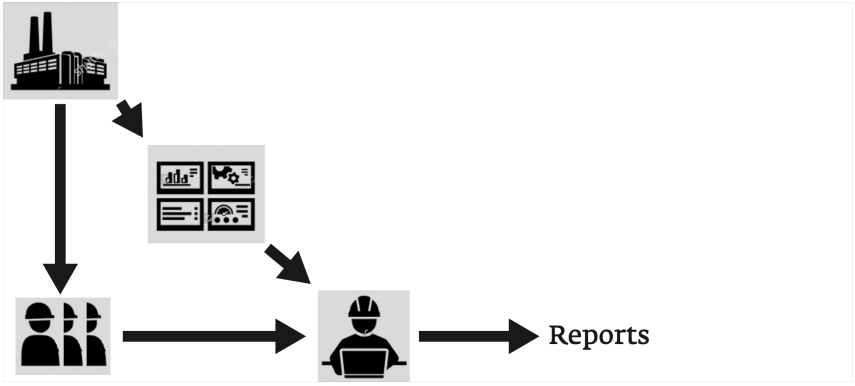
1. About Me
2. About Natural Language Processing (NLP)
3. Project Learn Artificial Intelligence (Learn-AI)
4. The Maintenance Corpus
5. Tailoring a controlled language out of the corpus
6. The Editor to help users to get used to the controlled natural language
7. Evaluation
8. Examples from Tailored Controlled Languages
9. Conclusion and Perspectives

Context: The Project (LEARN-IA)

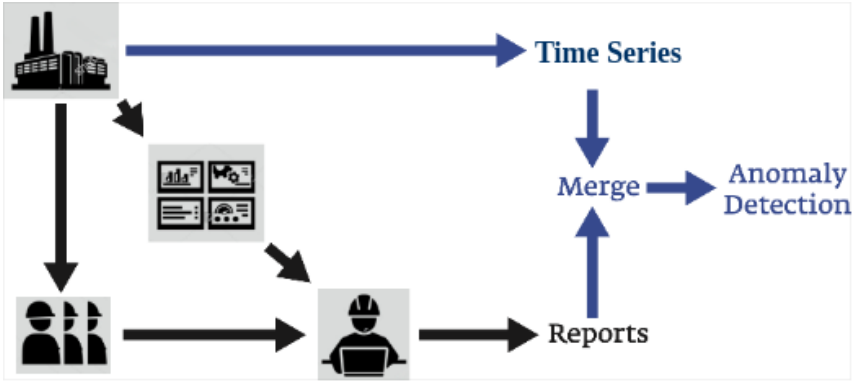
A thermal power plant in France, which burns charcoal or gas to produce electricity

Monitor, optimize its energy production, detect anomalies.

Goal 1 (this presentation): Study corpus of reports, optimize report writing process



Goal 2 (in progress): Merge data flows, detect anomalies



Outline

1. About Me
2. About Natural Language Processing (NLP)
3. Project Learn Artificial Intelligence (Learn-AI)
4. The Maintenance Corpus
5. Tailoring a controlled language out of the corpus
6. The Editor to help users to get used to the controlled natural language
7. Evaluation
8. Examples from Tailored Controlled Languages
9. Conclusion and Perspectives

Examples of reports

Rapport Jeu. 12 avr. 2018 06:00 Poste de matin

Poste E

Chef de poste : [redacted]
Tableauriste : [redacted]
PE : [redacted]
Présents : [redacted]
Absents : [redacted]

HSE

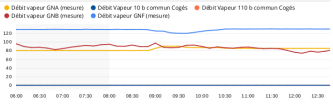
Actions Sécurité:
Formation du personnel: Formation T706

Emissions

	Moyenne
NOx Fumées GNF (mg/Nm3)	214

Production de vapeur

Débit vapeur produit par GNA, GNB, GNF, Cogénérations (t/h)



Indicateurs de performance

Écart à la consigne: pertes liées au rendement...

	Somme
Suivi température fum...	0,680
Suivi température fum...	0
Pertes tv O2 GNA (t)	0
Pertes tv O2 GNB (t)	0
Pertes tv O2 GNF (t)	0
Pertes tv T° Fumées GN...	0

Opti Chantier

	Moyenne
KPI TOT CHANTIER ( )	0,144

Écart à la consigne: cumul des tonnes de vape...

	Somme
Pertes tv MEA (0b et 10...	0
Pertes Optimisation Ch...	0,000
Pertes tv Rendement (...	0,680
Pertes mises à l'air non...	0

Commentaires sur performance énergétique :

Marche à 3 chaudières
Maintien parallèle avec T707 en régulation de puissance limitée à 8 MW

Exploitation courante GN :

Exploitation GN, réseau électrique (transfert sans coupures, NJS, alarmes dispatching) et manoeuvres spécifiques réalisées (démarrage/arrêt, isolement consignation pour travaux), etc.

GNF : Ramonage 13,17,1,3,2,8,9
Température air primaire: coefficients de 13400 m3/h à 13600 m3/h
Desox : Ras
Denox : Ras
GNB : En régulation de 10b
GNA : En fixe
Cogé L1 : Ras
Cogé L2 : Ras
Déminée :Régé chaine 1 / Contrôle silicemètre LM
Turboalternateurs : T707 : Vanne fermée sur mano sur prise d'impulsion vapeur 140b vanne gauche avant clapet Ne pas toucher (Voir PO)
T706: Essai de marche hors puissance des différents moteurs (Contrôle contact et retour de marche YSL)
Pompes alimentaires : V723 et V721 en service
Monocarbonatation :Augmentation du débit de lessive de 3.6 à 3.8 m3/h et débit de navettage sur Quench sur le seuil d'alarme à 10 m3/h
Réseaux électriques :
H2: réamorçage du réchauffeur 0b côté GNB

Avis de maintenance rédigés :

Points à traiter par équipe de jour :

Examples of comments

Commentaires sur performance énergétique :

GNA + GNB + GNF 100% gaz

Exploitation courante GN:

Exploitation GN, réseau électrique (transfert sans coupures, N/S, alarmes dispatching) et manœuvres spécifiques réalisées (démarrage/arrêt, isolement consignation pour travaux), etc.

GNF : _ Travaux de reprise bourrage ramoneuse 3 terminé/déconsigné. Essai à faire cet après-midi (si possible) en présence d[redacted]
_ Travaux sur transmetteur Q fumées (FIF702BD) GNF terminé. [redacted] doit repasser cet après-midi pour re-contrôler si la valeur lue est correcte. La calibration sera enlevé certainement demain matin.
_ Baisse du régulateur AC à 47.8%

Desox : A l'arrêt / en by-pass

Denox : En ligne

GNB : En service / en régulation de 10b

GNA : En service

Cogé L1 : _ Test sur VTL AF-103 en cours par équipe cogé, ne pas toucher
_ Déconsignation partie bâche alimentaire (vanne UV002...)
_ Cet après-midi, déconsigner la partie VVP/ACO et gaz CFM (BOX n°9 et 6)

Cogé L2 : _ Disponible en secours

Déminée : _ Régénération chaîne 1

Turboalternateurs : _ T702 à 1.2 MW, T706 à 3.3 MW et T707 à 7.8 MW

Pompes alimentaires : _ V721 en service et V722 en secours

Monocarbonatation : _ A l'arrêt. Vanne XVM002B ouverte en forcée pour circulation vers M001

Réseaux électriques : _ RAS

Divers:

_ RAS: appel UE, renvoi H2 de la salle 7 à 11h00. Une baisse est prévue à 19h ce jour
_ Travaux sur T708 en cours

Statistics of the corpus

- 2,280 maintenance reports, written in 8-hour intervals during 2 years (more has arrived since but has not yet been processed)
- 30,851 sentences in “telegraphic style”
- 138,140 words
- POS tag distribution:

ADJ	COMMUN NOUN	VERB	PROPER NOUN
7,137	43,034	21,911	18,305

Characteristics: High Spelling Variation

An example: “régénération” is the sixth most frequent word in the corpus.

régé	1,117	apocope
Régé	549	apocope
régénération	485	(**standard form**)
Régе	14	apocope & accent error
régе	12	apocope & accent error
rege	11	unaccented apocope
Regé	6	apocope & accent error
Rege	5	unaccented apocope
régés	5	apocoped plural
regé	4	apocope & accent error
Regénération	2	accent error
Régénaration, regenaration	1	spelling errors
régeneration, Régénèration	1	accent errors
regeneration, régéneration	1	accent errors

Characteristics: Low Morphological Variation

- Only three frequent (more than 50 occurrences) verb forms (not counting infinitives and participles):

third person singular at the present time	third person plurial at the present time	third person singular at the future time
2,638	56	106

Outline

1. About Me
2. About Natural Language Processing (NLP)
3. Project Learn Artificial Intelligence (Learn-AI)
4. The Maintenance Corpus
5. Tailoring a controlled language out of the corpus
6. The Editor to help users to get used to the controlled natural language
7. Evaluation
8. Examples from Tailored Controlled Languages
9. Conclusion and Perspectives

Tailor this Language Towards a Controlled Language

- Why:
  - Optimize report **parsing**
  - Improve information and knowledge extraction
  - Merge with time series data and detect anomalies
- Goal:
  - The designed controlled natural language must be a subset of the french language
  - The use of the controlled language should be almost natural
  - Switching to the controlled language should be progressive and painless
- How:
  - Normalizing vocabulary and simplify syntax

Tailor this Language Towards a Controlled Language

- Lexical normalization:
  - Vocabulary already quite limited
  - Define synonyms and abbreviations
  - Use a standard spell-checker
- Syntax:
  - Users (writers) and readers speak standard French in France
  - The designed controlled language should strive towards standard French in France
  - Keep (partly) the “telegraphic style” for brevity and to make the transition almost natural

The Tailoring Principle: Problem

- A natural language A in a standard form
  The set of syntax rules is noted as R(A)
- A corpus of “telegraphic-style” documents B in a given knowledge domain
  The set of syntax rules is noted as R(B)
- We want to obtain a controlled language that is
  1) based on corpora A and B (on syntax rules R(A) and R(B))
  2) has a relatively simple syntax described by a formal grammar
- Number of rules in the Controlled Natural Language:
  More rules: more expressive but more difficult to process
  Less rules: less natural but simpler to process

The Tailoring Principle: solution

- Lexical level: Control the vocabulary by using a configurable spell-checker.
- Syntax level:
  - Based on numbers of the most frequent syntax rules:
  - Choose a subset of rules in standard french corpus, $R(A) \supset R'(A)$
  - Choose a subset of rules in maintenance report corpus in french, $R(B) \supset R'(B)$
  - Combine R'(A) and R'(B) to obtain the rules set for controlled language
- The **Tailoring Principle**:
  - (1) By changing the proportion of rules from A and rules from B\A we get a balance between standard language and telegraphic style.
  - (2) By reducing the global number of rules we can strive towards formality and away from naturalness.

An Appropriate Standard French Corpus

- A corpus of best-seller novels
- Written in everyday, semi-formal, simple French
- Copy-edited and guaranteed error-free
- 11 novels processed = 48,693 non-elliptic sentences (neither titles nor ending with ellipsis) = 650,847 words

Corpora processing

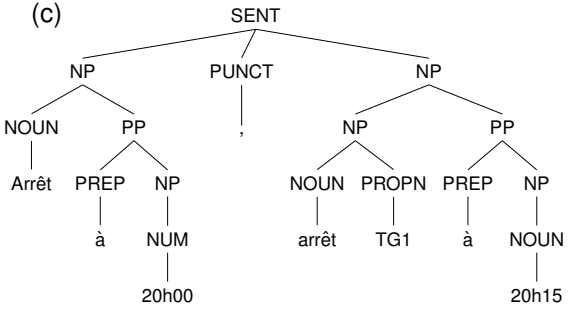
- For the report corpus:
  - Replaced all numerals and physical values with a single NUM tag and all device names and abbreviations with a PROPN tag
- For both corpora:
  - Use the Stanford CoreNLP parser in order to obtain Phrase-Structure Grammar syntax trees
  - Removed lexical leaves
  - Extracted all production rules

An example of this process

(a) Arrêt à 20h00, arrêt TG1 à 20h15

(b) Arrêt à NUM , arrêt PROPN à NUM

(c)



(d)

SENT	→	NP	PUNCT	NP
NP	→	NOUN	PP	
PP	→	PREP	NP	
NP	→	NUM		
NP	→	NP	PP	
NP	→	NOUN	PROPN	
PP	→	PREP	NP	
NP	→	NOUN		

Some Comparative Facts

	Report corpus (138,140)	Novel corpus (650,847)
# distinct rules	8,583	30,930
most frequent rule	PP $\rightarrow$ PREP NP	PP $\rightarrow$ PREP NP
frequency of most frequent rule	18,584	48,573
hapaxes rules (Zipf tail)	5,662 (66%)	23,203 (75%)
frequency of SENT $\rightarrow$ PP	263	0
rank of NP $\rightarrow$ DET NOUN	11th	2nd

Outline

1. About Me
2. About Natural Language Processing (NLP)
3. Project Learn Artificial Intelligence (Learn-AI)
4. The Maintenance Corpus
5. Tailoring a controlled language out of the corpus
6. The Editor to help users to get used to the controlled natural language
7. Evaluation
8. Examples from Tailored Controlled Languages
9. Conclusion and Perspectives

Train Tailored Language Users Using an editor

- In our context, users may be under stress.
- They should be aware of the result of the parsing operation, but not forced to obtain a successful parse.
- Showing syntax trees is not an option, but one could show segmentation in parsed sentences.
- It is natural to start parsing from the left and go as far as possible.
- We call this approach **Left-Right Maximal Segmentation**.

Example of Left-Right Maximal Segmentation

Language generated by

- a language with ≥ 20 / ≥ 10 frequencies
- the sentence: “en bouteille, voir schéma”

  en

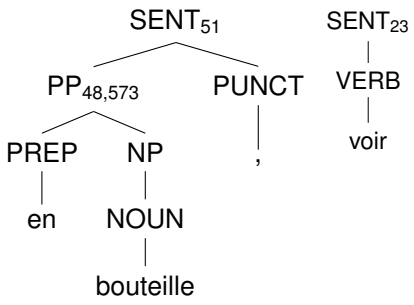
  [en bouteille]

  [en bouteille,]

  [en bouteille,] [voir]

  [en bouteille,] [voir] schéma

based on the trees:

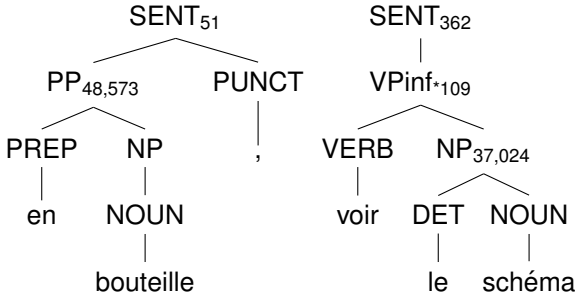


Example of Left-Right Maximal Segmentation

- To correct this sentence, the author may simply attempt to add the determiner in front of the noun:

  [en bouteille,] [voir le schéma]

- in which case the complete sentence is parsed, with the following trees:



Outline

1. About Me
2. About Natural Language Processing (NLP)
3. Project Learn Artificial Intelligence (Learn-AI)
4. The Maintenance Corpus
5. Tailoring a controlled language out of the corpus
6. The Editor to help users to get used to the controlled natural language
7. Evaluation
8. Examples from Tailored Controlled Languages
9. Conclusion and Perspectives

Evaluation

- The system has not yet been installed in the power plant, so we cannot evaluate the editor.
- But we can evaluate the coverage of the existing corpus with respect to various tailored languages:

$\mathfrak{M}$ \ $\mathfrak{S}$	≥ 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

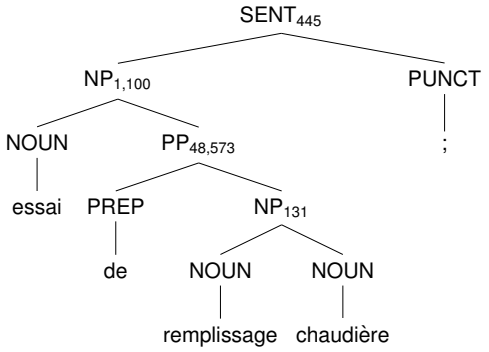
Outline

1. About Me
2. About Natural Language Processing (NLP)
3. Project Learn Artificial Intelligence (Learn-AI)
4. The Maintenance Corpus
5. Tailoring a controlled language out of the corpus
6. The Editor to help users to get used to the controlled natural language
7. Evaluation
8. Examples from Tailored Controlled Languages
9. Conclusion and Perspectives

Rules $\geq$ 50 in novels corpus

Language generated by

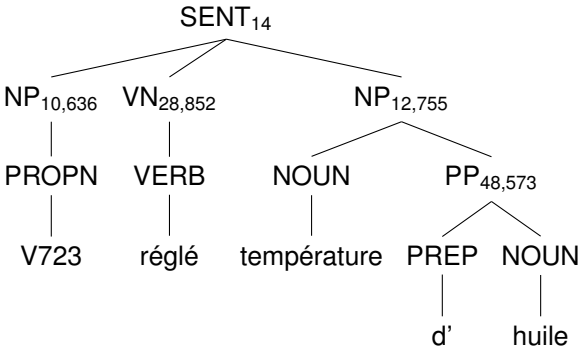
- Rules of frequency $\geq$ 50 in the novel corpus
- No rule from the report corpus
- Couverage of report corpus: 13.89%



Rules $\geq$ 10 in the novels corpus

Language generated by

- Rules of frequency $\geq$ 10 in the novel corpus
- No rule from the report corpus
- Couverage of report corpus: 28.72%

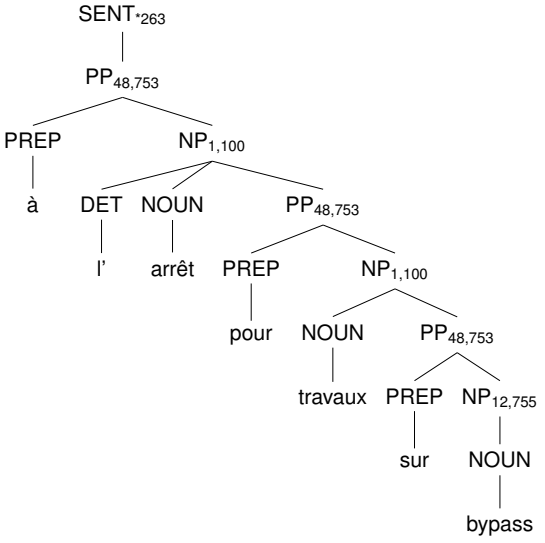


Frequency $\geq$ 50 in the novels corpus and frequency $\geq$ 10 in the reports corpus

Language generated by

- Rules of frequency $\geq$ 50 in novel corpus
- Rules of frequency $\geq$ 10 in report corpus
- Couverage of report corpus: 59.95%

(Number = frequency in novel corpus, *number = frequency in report corpus)



Outline

1. About Me
2. About Natural Language Processing (NLP)
3. Project Learn Artificial Intelligence (Learn-AI)
4. The Maintenance Corpus
5. Tailoring a controlled language out of the corpus
6. The Editor to help users to get used to the controlled natural language
7. Evaluation
8. Examples from Tailored Controlled Languages
9. Conclusion and Perspectives

Conclusion on Controlled natural languages and then for Anomaly Detection

- By selecting syntax rules among the most frequent ones in a corpus of “standard” French and in the existing “telegraphic-style” report corpus, we obtain *nested tailored controlled languages* that couver an optimal controlled language.
- By using a text editor we can help users to write sentences of the tailored language.
- Vocabulary and syntax become part of the feature space of the anomaly detection algorithm.

Questions & Discussions

The Learning Curve

- Let T, N be fixed sets of terminals and non-terminals, and S an initial symbol.
- For a set of production rules R, 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 $L(R') \subset L(R)$.
- Therefore by allowing only a subset of production rules we obtain a sub-language.
- By progressively reducing the number of allowed rules, we obtain a *sequence of nested tailored languages*

$$L(R_1) \supset L(R_2) \supset \cdots \supset L(R_\infty).$$

The progression can be standard, or depend on the individual user's progress, language characteristics, etc.

Where do I Get an Appropriate French Corpus?

- The FTB Corpus? No, too formal.
- A corpus of user forums or blogs? Not clean enough.
- A corpus of technical documents? Hard to find and often too formal.
- Proust's *Recherche du temps perdu*? Too long sentences.
- Simenon's complete detective stories? Too much slang (= poor syntax).
- Which corpus is simple, unpretentious, thoroughly copy-edited and easy to find in large quantities?

References

- Classification of Controlled Natural Languages
  - Tobias Kuhn. 2014. A survey and classification of controlled natural languages. Computational Linguistics, 40:121–170.
- Examples of existing Controlled Natural Lanugages:
  - **ASD-STE:** ASD (AeroSpace and Defence Industries Association of Europe), 2013. Simplified Technical English. Specification ASD-STE100, Issue 6.
  - **SLANG:** Rychtyckyj, Nestor. 2002. An assessment of machine translation for vehicle assembly process planning at Ford motor company. In Proceedings of AMTA2002, number 2499 in LNAI, pages 207–215. Springer.
- Controlled Natural Language Editor
  - Krasimir Angelov and Michal Boleslav Mechura. 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.
- Barbara Partee. 1995. Lexical semantics and compositionality. In An Invitation to Cognitive Science: Language, volume 1, pages 311–360. MIT Press.
- Horacio Saggion. 2017. Automatic Text Simplification. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.