Natural langage processing can support clear writing: the example of the AMesure platform

Thomas François

UCLouvain

CENTAL, IL&C

ELRC – Simplify Language

September, 24th 2021

Obstacles to clear writing

Administrative texts remain difficult to read for more than half of the citizens (Kimble, 1992).

Why?

1. **Need for prestige**: Deppert (1997) compares the reception of specialized vs. non-specialized readers of original and simplified texts.
   - simplified texts: better understood + more interesting
   - original texts: writer perceived as more prestigious!
2. **Need for assistance:** Simple writing guidelines are available, but underused (Nord, 2018):
   - general or vague principles, divergent positions, limited diffusion, etc.

Going deeper into writers' practices

Müller, Clerc and François (2021, *Discourse and Writing*).

Professional writers

- Cross-sectional survey on 55 writers
- 35 questions about practices
- 9 excerpts to simplify

Müller and François *(2022, in press)*

Functional writers

- Cross-sectional survey on 51 writers
- 35 questions about practices
- 9 excerpts to simplify

Going deeper into writers' practices

Some interesting results compared:

Category	Assertion	% of writers that "agree" or "strongly agree"	
		Professional	Functional
Plain language criticism	Plain language is less accurate	9%	23%
	Plain language loses the subtleties of language	13%	41%
Environmental limits	Good training in clear writing	58%	51%
	Pressure from superiors to write in a certain way	42%	23%
	Supported by their colleagues	69%	47%
Types of aid used	Dictionaries (usual, synonyms)	100%	78%
	Proofreading by another person	82%	92%
	Plain language guidelines	64%	43%

Going deeper into writers' practices

Analysis of the simplifications done and writers' characteristics

Table 6. Spearman correlations between simplification levels and writers' characteristics (* = significant at the 0.05 level; ** = significant at the 0.01 level)

	Years of experience	Importance of simplification in work time	Good training	Use of guidelines	Type of guide used	Confidence in their practice
Lexical features	-0.126	-0.019	-0.081	**0.417****	0.298	-0.225
Syntactic features	0.163	**0.432****	0.215	0.142	-0.126	0.057
Structural features	0.170	0.149	0.116	0.226	0.229	0.130
Relational aspects	-0.082	**0.447****	-0.004	**0.391***	**0.362***	0.127
Visual aspects	-0.221	0.165	-0.021	0.276	**0.465****	-0.057

Müller, Clerc and François (2021:67, *Discourse and Writing*).

Going deeper into writers' practices

Analysis of the simplifications done and writers' characteristics

- Experience (practice) is sig. for functional writers and for pro. (syntactic & relational)
- Training does not make any difference for both types
- Plain language guidelines help more functional writers

		Niveau exp.	Années exp.	Bonne formation	Utilisation d'un guide ou non	Confiance dans la pratique
Aspects lexicaux	Corrélation Spearman	**0,487****	0,271	-0,083	-0,212	0,299
Aspects syntaxiques	Corrélation Spearman	**0,546****	0,342	0,167	**-0,465****	0,294
Structure du texte	Corrélation Spearman	**0,574****	0,295	0,164	**-0,438***	**0,45***
Aspects relationnels	Corrélation Spearman	**0,657****	0,302	0,175	**-0,507****	**0,419***
Aspects visuels	Corrélation Spearman	**0,417***	**0,364***	0,144	**-0,363***	**0,368***

* significativité à 0,05 / ** significativité à 0,01

Müller and François (2022, *in press*).

Our proposal: AMesure (François et al. 2020)

AMesure aims to help writers to remember and apply simple reading guidelines, providing:

- A global readability score
  (readability formula, in A)
  [Francois et al, 2014]
- Assessment of several linguistic dimensions of the text (B)
- Highlighting complex phenomena in the text (C)
- Suggestions for simple writing for each sentence (D)

Previous work

Writing studies:

- Clear writing guidelines [Gouvernement du Québec, 2006, Ministère de la Communauté française de Belgique, 2010, Union européenne, 2011, Cutts, 2020]
- Studies on clear writing [Kimble, 1992, Labasse, 1999, Labasse, 2001, Desbiens, 2008, Clerc, 2009, Adler, 2012]

Automatic text simplification:

- Various ATS approaches (rule-based, MT, NMT)
  [Shardlow, 2014, Siddharthan, 2014, Saggion, 2017]
- Previous clear writing platforms
  [Scarton et al., 2010, Lee et al., 2016, Falkenjack et al., 2017, Yimam and Biemann, 2018]

A. Readability formula

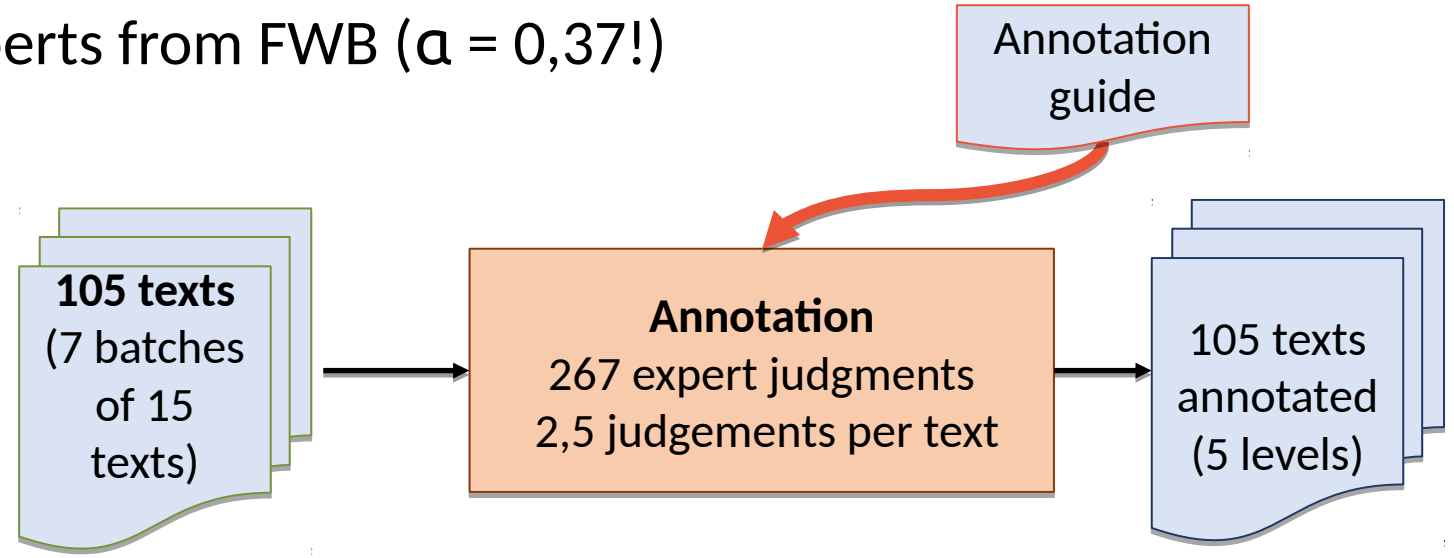
Annotation of the training corpus:

1) Collecting 115 authentic administrative texts

2) 10 texts read by subjects → ranked by reading time and Kandel and Moles (1958)

   1)output : **annotation guide** + scale with 5 levels of difficulty

3) Annotation process by 18 experts from FWB (α = 0,37!)



A. Readability formula

Creating the formula:

1) Selecting the best predictors

2) Training a ML model (SVM)

Accuracy = 58 %

Adj. Acc. = 91 %

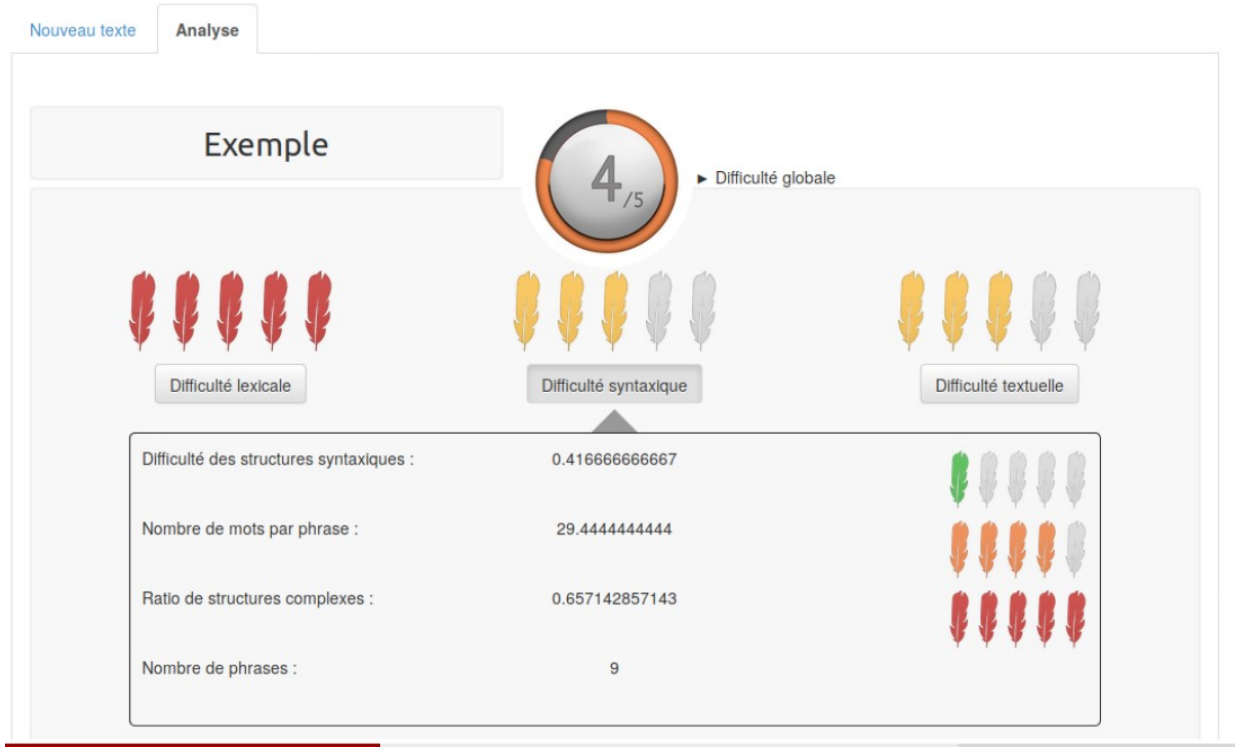
3) Predict over a 5-point difficulty scale

Variable description	p
Unigram model based on frequencies	-0,32
Median of the frequencies of verbs in the text	-0,47
Proportion of absent words from Gougenheim 8000	0,44
Type-Token ratio (lemmas)	-0,21
Proportion of words longer than 8 letters	0,40
Average cumulated freq. of orthographic neighbours	0,50
# words / # sentences	0,64
# past participle verbs / # verbs	0,46
# conjunctions / # pronouns	0,54
# P1 and P2 pronouns / # words	-0,42

B. Difficulty ID of the text

We have selected 11 readability yardsticks

- 5 lexical, 4 syntactic, and 2 textual



C. Detecting complex phenomena in administrative texts

Currently detected:

- Subordinated clauses:
  - relative clauses
  - object clause (fr. *complétive*)
  - adverbial clause
- Passive sentence
- Brackets
- Technical terms (list-based)
- Abbreviations (list-based and rule-based)
- Complex words (frequency-based)

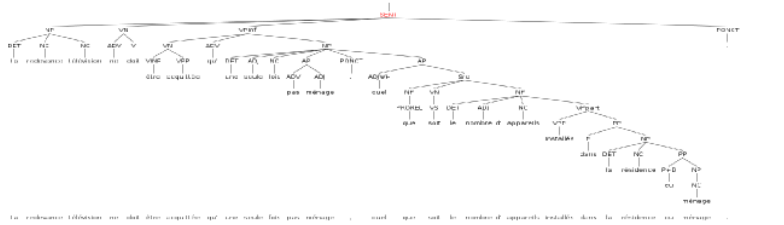
The detection of the syntactic structures

Implementation based on [Brouwers et al., 2014]

Sentences to analyse

En Région wallonne, une taxe annuelle d'un montant de 100 € doit être payée lorsque l'on détient un appareil de télévision, quel que soit l'usage qui en est fait.

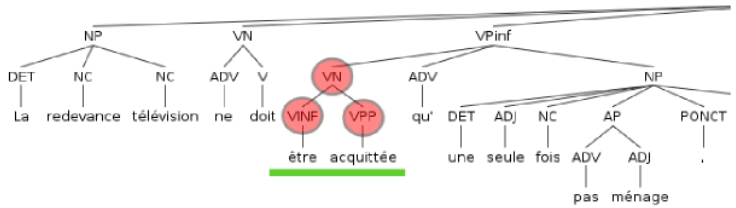
Syntactic parsing (Berkeley Parser)



Defining rules (based on a corpus)

```
"VN < (V|VINF|VPP "+etre()+" $.. (VPP "+notVIntransitif()+"))";
```

Applying regular expressions (via Tregex)



Evaluation

Test data = 24 administrative texts (637 clauses, 356 passives, 73 abbreviations)

Phenomena	R	P	F1	κ
Passive clauses	0.92	0.92	0.92	0.92
Subordinated clauses (all)	0.84	0.87	0.85	0.47
Relative clauses	0.83	0.88	0.86	/
Object clauses	0.56	0.42	0.48	/
Adverbial clauses	0.78	0.83	0.8	/
Abbreviations	0.73	0.9	0.8	0.97
Total (macro-average)	0.83	0.9	0.86	0.79

Table – Recall (R), precision (P), F1, percentage of agreement and Fleiss' κ scores for the five phenomena detected in the platform.

D. Generating the advice

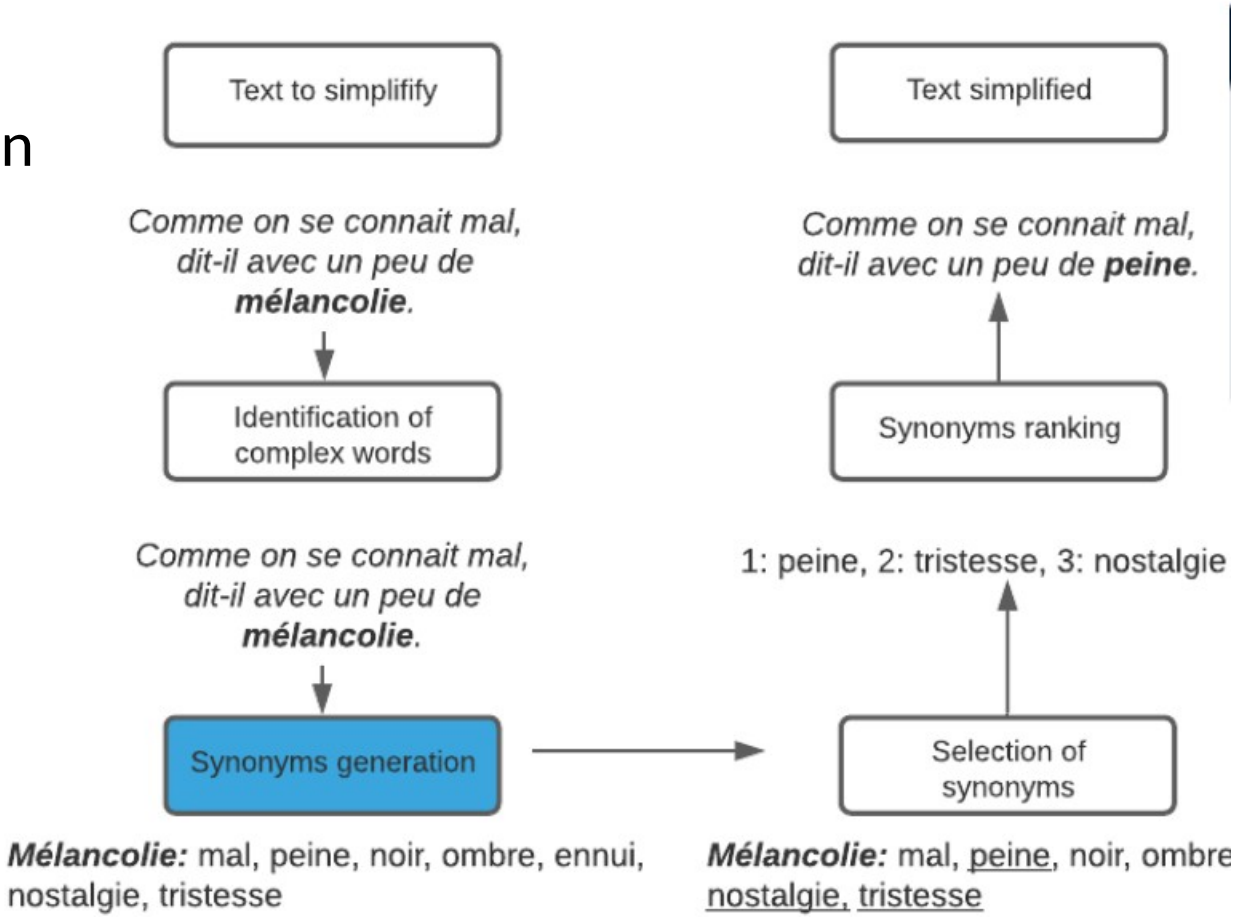
- Theoretical reference = simple writing guides from administrations
- 7 cases has been implemented so far

Problem	Condition
number of nested syntactic structures	≥ 3
total number of clauses	> 3
length of the sentence	> 15 words
length of the longest nested clause	> 10 words
length of text between brackets	> 10 words
number of subordinated clauses	≥ 3

Current research : generating simpler synonyms

Rolin et *al.*, *2021* (RANLP)

- 4 common steps for lexical substitution
- 2 main approaches for generation:
  - Lexicon (ResyF, Bilami et al., 2018)
  - Word embeddings (FastText, BERT)
- Difficulty ranking :
  - SVM model (François et al., 2016)



Quick demo

Conclusions

AMesure: 1st platform using NLP to suppor clear writing of French administrative texts

- Freely available (supported by FWB).
- Combines NLP technologies and clear writing studies

Perspectives:

- More tests with professional and functional writers
- Enriching the range of linguistic phenomenon detected and the advice set (Ph.D. thesis of Mrs. Müller)

FÉDÉRATION WALLONIE-BRUXELLES

Thank you for your attention