HECTOR: A Hybrid TExt SimplifiCation TOol for Raw Texts in French

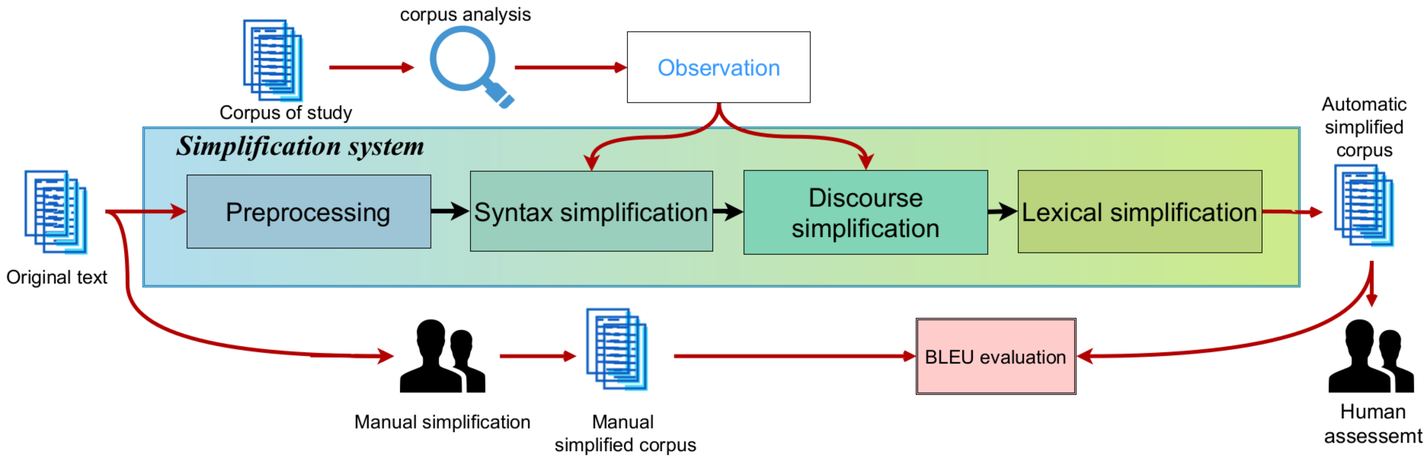
AMALIA TODIRASCU[1], RODRIGO WILKENS[2], EVA ROLIN[2], THOMAS FRANÇOIS[2], DELPHINE BERNHARD[1], NÚRIA GALA[3]

[1] Université de Strasbourg, LiLPa UR 1339, F-67000 Strasbourg, France
[2] CENTAL, Université catholique de Louvain, Belgium
[3] Aix-Marseille Université, Laboratoire Parole et Langage, LPL CNRS (UMR 7309), France

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR

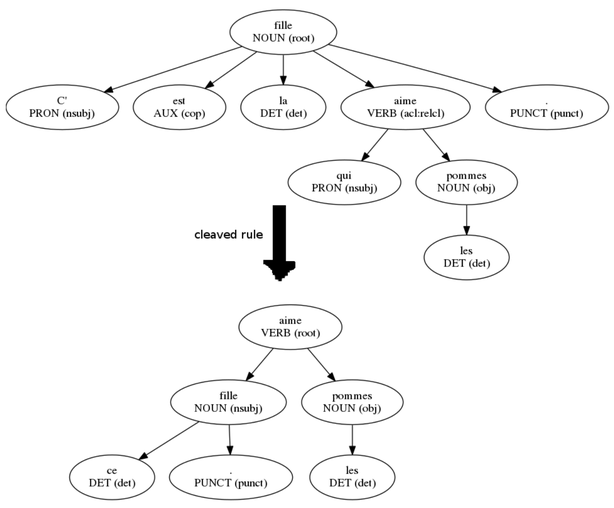
GOALS AND CONTEXT

- Few systems/resources are available for French: MUSS (Martin et al, 2020), parallel corpus for medical texts (Cardon et Grabar, 2018)
- ALECTOR (funded by ANR : ANR-16-CE28-0005): Building a hybrid, modular, interpretable, automatic text simplification system for French for people with dyslexia
  - Syntactic and discourse level: rule-based approach
  - Lexical level: a combination of statistical methods and word embeddings



SYNTACTIC SIMPLIFICATION

- 10 rules described in a specific langage, based on SemGrex (Levy and Andrew, 2006) and Tsurgeon (Wilkens et al, 2020a)
- Specific guidelines (Gala et al, 2020):
  - deleting secondary information
  - splitting complex phrases
  - sentence structure adjustments



DISCOURSE SIMPLIFICATION

- Modifying coreference chains: at least 3 referring expressions (RE) such as proper nouns, noun phrases and pronouns (Schnedecker, 1997)
- 3 categories of rules, based on accessibility theory (Ariel, 1990) and CoFR (automatic coreference annotation) (Wilkens et al, 2020b)
  - Replace new or repeated entities: pronouns→ explicit referent
    - i. **Le loup** va et vient. **Il** fixe le garçon. 'The wolf goes back and forth. It stares at the boy.'→ **Le loup** va et vient. **Le loup** fixe le garçon. 'The wolf goes back and forth. The wolf stares at the boy.'
  - Specify entities: high accessible RE → less accessible mentions
    - ii. **ce** hérisson 'this hedgehog' → **le** hérisson 'the hedgehog'.
  - Make NP more accessible: Possessive NPs are replaced by their explicit referent detected by CoFR (Wilkens et al, 2020b)
    - iii. **Sa** grand-mère 'Her grandmother' → **la grand-mère du Chaperon rouge** 'The grandmother of Little Red Riding Hood.'.

LEXICAL SIMPLIFICATION

Complex word identification:
(1) frequency < 5 per million, based on Lexique3 (New et al., 2007)
(2) it meets at least one of the criteria: word length > 7 ; absent from Manulex (Lété et al., 2004); an etymological letter is detected (Gala and Ziegler 2016); the distance between phonemic and orthographic forms > 2.

Substitution generation:
Generating synonyms with FastText (Bojanowski et al, 2017)

Substitution selection:
Disambiguating synonyms (closest semantic neighbours, POS filter)

Substitution ranking:
Using frequency from Lexique3 (New, 2006) to select simplest synonyms (high frequence words)

EVALUATION METHODS

- Evaluation corpus: original and manually simplified sentences from the ALECTOR corpus (Gala et al, 2020)
- Automatic evaluation: BLEU (Papineni et al., 2002)

Corpus	Genre	Doc. level	Sent. level
CM1 (4th)	SCI	.54 (.11)	.51 (.24)
	LIT	.64 (.10)	.60 (.28)
CE1 (2nd)	SCI	**.76 (.06)**	**.78 (.24)**
	LIT	.73 (.03)	.64 (.29)
	-	.67 (.12)	.62 (.28)

- Human evaluation
  - 3 criteria (meaning preservation, fluency, simplicity): Likert scale (1 – 5)
  - 3 coders per level (syntax, discourse) and 2 coders for lexical level
  - Inter-coder agreement (Krippendorf α)

	Syntax	Discourse	Lexical
Meaning preservation	0.58	0.265	0.455
Fluency	0.738	0.632	0.457
Simplicity	0.485	0.29	0.369

HUMAN EVALUATION RESULTS

- Meaning preservation (≥3.5)

	Syntax	Discourse	Lexical
CE1 Litt (2nd)	80.95%	70.00%	66.67%
CE1 Sci (2nd)	50.00%	93.33%	46.00%
CM1 Litt (4th)	73.68%	88.23%	38.88%
CM1 Sci (4th)	77.77%	88.23%	40.62%
Total	70.60%	83.34%	48.04%

- Fluency (≥ 3.5)

	Syntax	Discourse	Lexical
CE1 Litt (2nd)	91.89%	87.50%	86.60%
CE1 Sci (2nd)	69.23%	73.33%	73.73%
CM1 Litt (4th)	90.00%	100%	94.44%
CM1 Sci (4th)	89.65%	68.42%	93.75%
Total	85.19%	82.31%	87.13%

- Simplicity (≥3.5)

	Syntax	Discourse	Lexical
CE1 Litt (2nd)	84.21%	70.00%	86.00%
CE1 Sci (2nd)	50.00%	27.27%	46.66%
CM1 Litt (4th)	84.21%	86.66%	66.66%
CM1 Sci (4th)	82.35%	23.53%	40.62%
Total	79.36%	50.90%	59.29%

Error analysis

- Syntax simplification: wrong word order
  - a. Aussitôt les villageois se précipitent vers l'arbre et se gorgent de ces fruits sublimes./'Immediately the villagers rush to the tree and stuff themselves with the sublime fruits.' → *Les villageois **Aussitôt** se précipitent vers l'arbre. ***De ces fruits sublimes** Les villageois se gorgent. /'The villagers immediately rush to the tree. *with the sublime fruits the villagers stuff themselves .'
- Discourse simplification: changing determiners -> changing referent
  - b. Les algues produisent plus de la moitié de l'oxygène de notre air./'Algae produce more than half of the oxygen in our air.' → Les algues produisent plus de la moitié de l'oxygène **du** air. /'Algae produce more than half of the oxygen in the air.'
- Lexical simplification: MWE or terms
  - c. Le gypse est une roche tendre./'Gypsum is a soft stone' → ***Un plâtre** est une roche tendre./'A plaster is a soft stone)'