Rated Lexicon for the Simplification of Medical Texts

Anaïs Koptient, Natalia Grabar

UMR 8163 STL CNRS, Université de Lille

anais.koptient.etu@univ-lille.fr natalia.grabar@univ-lille.fr

Presenter

- Anaïs Koptient, 2nd-year PhD student at the University of Lille, France
- PhD topic : Fine-grained simplification of technical documents

Automatic text simplification

- make a text more understandable for a specific group of persons (children, people with pathologies, foreigners, etc.)
- very little work made on specialized texts (like medical texts) and in French
- several levels of simplification :
  - **lexical simplification :** difficult words replaced by simpler equivalent,
  - **syntactic simplification :** syntactically complex sentences divided into simpler sentences,
  - **semantic simplification :** information is reorganized,
  - **pragmatic simplification :** structure of the text is modified.

Automatic text simplification

- 3 approaches :
  - approaches based on distributional probabilities (word embeddings) [Glavas and Stajner, 2015, Kim et al., 2016],
  - approaches based on automatic translation [Zhao et al., 2010, Wubben et al., 2012, Sennrich et al., 2016, Nisioi et al., 2017],
  - approaches based on rules [Carroll et al., 1999, Bautista et al., 2009, De Belder et al., 2010].
- ==> **need for resources**

Purpose of our work

- build a lexical resource with French medical terms :
  1. identify lexical equivalents for technical medical terms,
  2. assign a readability score to technical terms and their equivalents.

Outline

- Identification of lexical equivalents for technical terms
- Approaches for rating the lexicon and evaluation
- Conclusion and future works

2. Identification of lexical equivalents for technical terms

- Corpora used
  1. CLEAR corpus [Grabar and Cardon, 2018]
  2. forum *masante.net*
- Methods
  1. extraction of equivalents from parallel aligned sentences
  2. definitions of technical terms
  3. reformulations of technical terms
  4. word morphology
  5. expansion of abbreviations
  6. exploitation of an online medical dictionary

2. Identification of lexical equivalents for technical terms

Corpora used

CLEAR Corpus

- Corpus composed of medical comparable texts differenciated by their technicality and difficulty
- 16,313 pairs of texts
- 3 sub-corpora :

Title of the corpus	Technical part	Simple part
Drug leaflets [1]	Drug leaflets created for medical doctors	Durg leaflets found in drug boxes
Abstracts of systematic review (http ://www.cochranelibrary.com/)	Technical abstracts	Manually simplified versions of the technical abstracts
Encyclopedia articles	Medicine-related articles from French Wikipedia (https ://fr.wikipedia.org)	Correspondind articles from the French children version (https ://fr.vikidia.org)

1. http ://base-donnees-publique.medicaments.gouv.fr/

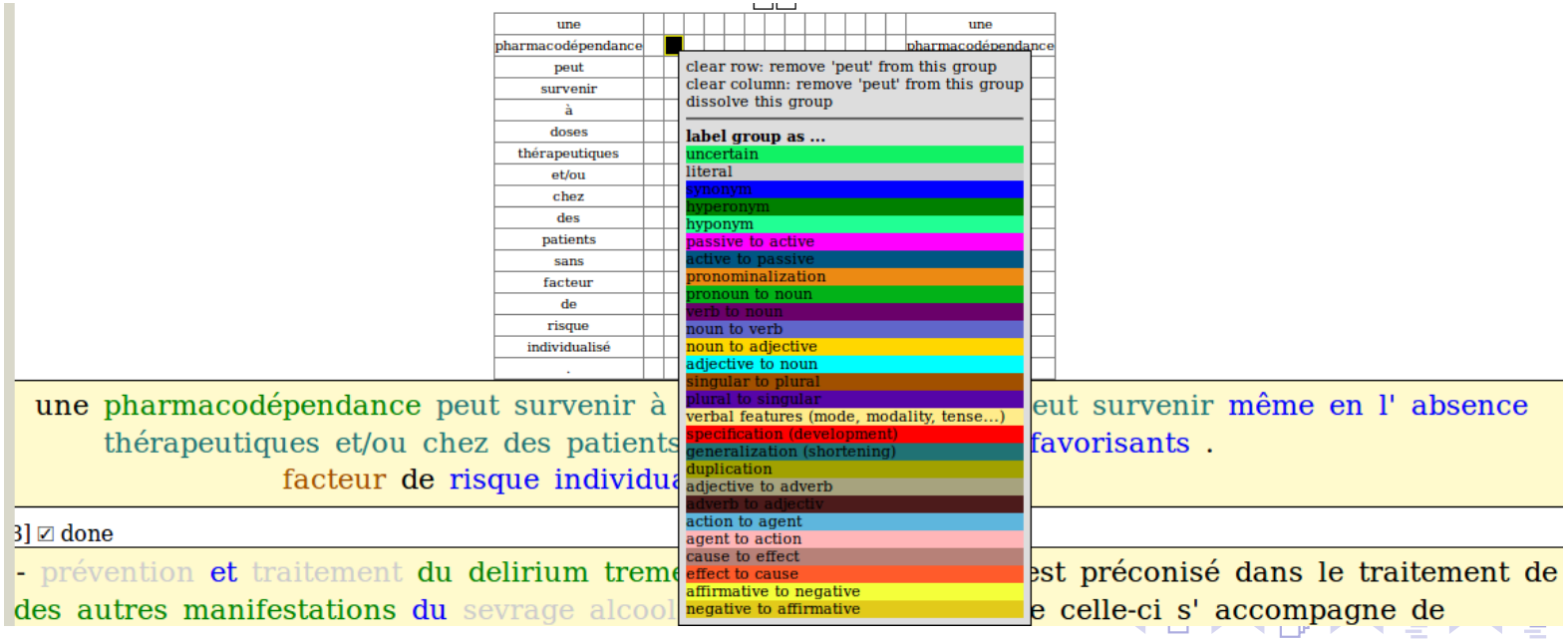
Forum from *Masante.net*

- Forum which provides answers from medical doctors to questions related to health
- 6,139 answers exploited

2. Identification of lexical equivalents for technical terms

Methods for indentification of lexical equivalents

1. Extraction of equivalents from parallel aligned sentences
   - Manually aligned sentences from CLEAR corpus
   - Manual annotation of the transformations observed during simplification [Koptient et al., 2019]
   - Extraction of equivalents corresponding to the *synonym* and *hyperonym* annotations => **626** pairs technical term/simpler equivalent



2. Definitions of technical terms
   - Exploitation of context terms like *est un (is a)* or *défini comme (defined as)* :
     - *L'**angiographie** est une **technique d'imagerie médicale portant sur les vaisseaux sanguins qui ne sont pas visibles sur des radiographies standards**. (**Angiography** is a **medical imaging technic for blood vessels which are not visible with standard imaging**.)*
   - => **1,028** definitions

3. Reformulations of technical terms
   - Exploitation of reformulations between brackets :
     - *Vous avez effectivement une hématurie (trop de globules rouges dans vos urines). (Indeed, you have hematuria (too much red blood cells in urine).)*
   - Exploitation of reformulation markers (*c'est-à-dire (that is (to say))*, *autrement dit (in other words)*, *l'équivalent (the equivalent)* or *encore appelé (also called)*) :
     - *La prise de poids est normale dans la périménopause, c'est à dire la période qui entoure la ménopause. (Weight gain is expected during perimenopause, that is the period which surrounds the menopause.)*
   - => **7,959** pairs technical term/simpler equivalent

4. Word morphology
   - Combination :
     - Set of Latin and Greek affixes and their semantics
     - Combination of each prefix with each suffix :
       - *angio (blood vessel)* + *logy (study of)* = **angiology|study of blood vessels**
     - => **1,939** pairs technical term/simpler equivalent
   - Transformation into morphological bases :
     - Terms analyzed with Dérif [Namer, 2009] to transform them into morphological bases :
       - *myocardique (myocardial)* => *myo (muscle)* and *carde (heart)*
     - Search into the corpus to find syntactic groups that contain the meaning of the bases :
       - found sequences that contain *heart muscle* meaning *muscle du coeur* in French

5. Expansion of abbreviations
   - Extraction of expanded forms of abbreviations using adapted version of published algorithm [Schwartz and Hearst, 2003]
   - Two kinds of structures extracted :
     - expanded form (abbreviation) :
       - *On l'appelle aussi liquide cérébro-spinal (LCR). (It is also called cerebrospinal fluid (CSF).)*
     - abbreviation (expanded form) :
       - *Le finastéride a été retrouvé dans le LCR (liquide céphalo-rachidien) (Finasteride has been found in CSF (cerebrospinal fluid).)*
   - => **8,148** pairs abbreviations/expanded form

6. Exploitation of an Online Medical Dictionary
   - Exploitation of the online lexicon
     https ://www.cancer.be/lexique : for each technical term, the first sentence is extracted as simpler equivalent
   - => **1,165** pairs technical terms/simpler equivalent

2. Identification of lexical equivalents for technical terms

Results

Methods	*# extractions*	*Precision*
Parallel sentences	626	100
Definitions	1,028	68
Reformulation	7,959	60
Morphological analysis	1,128	86
Morphological affixes and roots	1,939	13
Abbreviations	8,148	94
Online resources	1,165	100
English medical terms [Zeng et al., 2005]	11,641	–
English medical abbreviations [Schwartz and Hearst, 2003]	785	95
French medical terms [Deléger and Zweigenbaum, 2008]	147	67
French medical terms [Cartoni and Deléger, 2011]	109	66

3. Computing the readability of technical terms and their equivalents

Purpose :

1. Assign readability scores to the lexicon,
2. Verify if the paraphrases are easier than technical terms,
3. If necessary, switch the paraphrase and the technical term,
4. Provide indication on simplicity of terms and their equivalents.

3. Computing the readability of technical terms and their equivalents

Two types of readability forumulas :

1. Linear regression readability formulas,
2. Computational readability models.

3. Computing the readability of technical terms and their equivalents

Linear regression readability formulas

Dale index [Dale and Chall, 1948]

- $Dale = 0.15x1 + 0.04x2$
- $x1$ = percentage of words missing from the basic vocabulary
- $x2$ = average number of words per sentence
- the more the index is high, the less the text is readable

3. Computing the readability of technical terms and their equivalents

Linear regression readability formulas

Kandel index [Kandel and Moles, 1958]

- $Kandel = 207 - (1.015 * ASL) - (73.6 * ASW)$
- $ASL =$ average number of words in each sentence
- $ASW =$ average number of syllables
- index value between 0 and 100 : 0 to 30 = difficult to understand ; starting from 70 = easy to understand

3. Computing the readability of technical terms and their equivalents

Linear regression readability formulas

Mesnager index [Mesnager, 1989]

- $Mesnager = (1/2 * AC) + (1/3 * P)$
- $AC =$ percentage of words missing from basic vocabulary
- $P =$ average number of words in sentences
- index value between 6 (easy text) and 25 (difficult text)

3. Computing the readability of technical terms and their equivalents

Linear regression readability formulas

Sitbon index [Sitbon et al., 2010]

- $Sitbon = 1.12 * ADV - 0.69 * CON + 6.48 * cohesion + 15.58$
- $ADV =$ number of adverbs
- $CON =$ number of conjunctions
- $cohesion =$ number of phonemes divided by number of letters

Smith index [Smith, 1961]

- $L = -6.49 + 1.56WL + 0.19SL$
- WL = number of letters
- SL = number of words

3. Computing the readability of technical terms and their equivalents

Computational readability models

Use of descriptors issued from existing typology [Gala et al., 2014] :

- number of letters
- number of phonemes
- number of syllables
- cohesion between phonemes and spelling
- frequency
- presence in the Catach list [Catach et al., 1984]
- syllable components

3. Computing the readability of technical terms and their equivalents

Models

- Biclass model (*simple and difficult*)
- *Three classifiers as implemented by Scikit-Learn [Pedregosa et al., 2011] :*
  - *MultiLayer Perceptron,*
  - *Decision Tree,*
  - *Random Forest.*
- *The more the term or paraphrase is close to 0, the more difficult the term or paraphrase is*
- *Training of reference data with a 10-fold cross-validation :*

	Precision	*Recall*	*F-measure*
MLP	90.3	90.4	90.0
DT	88.7	89.0	88.6
RF	89.2	89.5	89.2

3. Computing the readability of technical terms and their equivalents

Results

- Technical terms and paraphrases rated for their readability by :
  - five readability indexes *Dale*, *Kandel*, *Mesnager*, *Sitbon* and *Smith*,
  - the proposed computational readability models.

3. Computing the readability of technical terms and equivalents

Results

Terms and their equivalents	*Dale*	*Kandel*	*Mesnager*	*Sitbon*	*Smith*	*MLP*	*DT*	*RF*
difficult	high	low	high	high	high	0	0	0
simple	low	high	low	low	low	1	1	1
comédon (comedo)	15.04	-235.615	66.33	22.06	4.62	0	0	0
point noir (blackhead)	0.08	102.77	0.66	21.34	0.91	1	1	1
vomissements (comiting)	15.04	-88.415	66.33	20.98	12.42	1	1	1
être malade (being sick)	0.08	65.98	0.66	20.76	1.69	1	1	1
lupus érythémateux disséminé (systemic lupus erythematosus)	15.12	-65.91	66.99	21.06	7.6	0.33	0.33	0.66
éruption faciale, douleur articulaire, anomalies musculaires, fièvre (facial eruption, articular pain, muscular abnormalies, fever)	15.4	16.91	69.3	21.11	5.082	0.67	0.67	0.67
condylomes acuminés (condylomata acuminata)	15.08	204.97	66.66	15.58	-6.11	0	0	0
verrues génitales (genital warts)	15.08	20.97	66.66	20.035	6.37	1	1	1
système endocrinien (endocrine system)	15.08	131.37	66.66	21.7	7.93	0.5	0.5	0.5
groupe de glandes et de cellules du corps fabriquant et libérant des hormones dans le sang, qui contrôlent de nombreuses fonctions comme la croissance, la reproduction, le sommeil, la faim et le métabolisme (group of glands and cells in the body that make and deliver hormones in blood, that control many functions such as growth, reproduction, sleep, hunger and metabolism)	9.97	73.7	49.26	18.05	8.00	0.67	0.67	0.5
alpha-foetoprotéine (afp) (alpha-foetoproteine (AFP))	15.08	-15.83	66.66	19.9	12.61	0	0	0
protéine normalement fabriquée par le placenta lors de la grossesse habituellement non présente dans le sang d'une femme en bonne santé qui n'est pas enceinte ou d'un homme en bonne santé (protein that is normally made by placenta during pregnancy, and usually missing in blood of healthy non-pregnant women or healthy men)	8.42	86.45	42.28	25.37	7.17	1	1	1
ostéologie (osteology)	15.04	-126.23	66.33	21.412	9.3	0	0	0
étude de l'os (study of bones)	7.66	94.57	34.32	19.11	-1.44	1	1	0.5
bézoards (bezoars)	15.04	-126.23	66.33	21.25	6.18	0	0	0
concrétions gastro-intestinales (gastrointestinal concretions)	15.08	204.94	66.66	21.41	17.29	0.5	0.5	0.5

3. Computing the readability of technical terms and equivalents

Results

- Sitbon index rather sensitive to long terms and paraphrases
- Overall, paraphrases are easier than technical terms
- For some pairs, both paraphrase and terchnical term are considered as understandable
- May be difficult to rate long paraphrases

4. Conclusion

- Creation of a lexicon for automatic text simplification : 11,272 pairs technical term/simpler equivalent
- Rating the lexicon for its readability with readability indexes and computational models

Références I

Bautista, S., Gervás, P., and Madrid, R. I. (2009).

Feasibility analysis for semi-automatic conversion of text to improve readability.
In *Int Conf on Inform and Comm Technology and Accessibility (ICTA)*, pages 33–40.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999).

Simplifying text for language-impaired readers.
In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway. Association for Computational Linguistics.

Cartoni, B. and Deléger, L. (2011).

Découverte de patrons paraphrastiques en corpus comparable : une approche basée sur les n-grammes.
In *Traitement Automatique des Langues Naturelles (TALN)*.

Catach, N., Jejcic, F., and de la Recherche Scientifique), E. H. C. N. (1984).

Les listes orthographiques de base du français (LOB) : les mots les plus fréquents et leurs formes fléchies les plus fréquentes.
Nathan, Paris, France.

Dale, E. and Chall, J. S. (1948).

A formula for predicting readability.
The Journal of Educational Research, 27(11-20) :37–54.

De Belder, J., Deschacht, K., and Moens, M.-F. (2010).

Lexical simplification.
In *ITEC*, pages 1–4.

Deléger, L. and Zweigenbaum, P. (2008).

Paraphrase acquisition from comparable medical corpora of specialized and lay texts.
In *Ann Symp Am Med Inform Assoc (AMIA)*, pages 146–50.

Références II

Gala, N., François, T., Bernhard, D., and Fairon, C. (2014).

A model to predict lexical complexity and to grade words (un modèle pour prédire la complexité lexicale et graduer les mots) [in French].

In *Proceedings of TALN 2014*, pages 91–102, Marseille, France.

Glavas, G. and Stajner, S. (2015).

Simplifying lexical simplification : Do we need simplified corpora ?

In *ACL-COLING*, pages 63–68.

Grabar, N. and Cardon, R. (2018).

Clear – simple corpus for medical French.

In *Workshop on Automatic Text Adaption (ATA)*, pages 1–11.

Kandel, L. and Moles, A. (1958).

Application de l'indice de flesch à la langue française.

The Journal of Educational Research, 21 :283–287.

Kim, Y.-S., Hullman, J., Burgess, M., and Adar, E. (2016).

SimpleScience : Lexical simplification of scientific terminology.

In *EMNLP*, pages 1–6.

Koptient, A., Cardon, R., and Grabar, N. (2019).

Simplification-induced transformations : typology and some characteristics.

In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 309–318, Florence, Italy. Association for Computational Linguistics.

Mesnager, J. (1989).

Lisibilité des textes pour enfants : un nouvel outil ?

Références III

Namer, F. (2009).

Morphologie, Lexique et TAL : l'analyseur DériF. TIC et Sciences cognitives.
Hermes Sciences Publishing, London.

Nisioi, S., Stajner, S., Ponzetto, S. P., and Dinu, L. P. (2017).

Exploring neural text simplification models.
In *Ann Meeting of the Assoc for Comp Linguistics*, pages 85–91.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011).

Scikit-learn : Machine learning in Python.
Journal of Machine Learning Research, 12 :2825–2830.

Schwartz, A. S. and Hearst, M. A. (2003).

A simple algorithm for identifying abbreviation definitions in biomedical text.
In *Pacific Symposium on Biocomputing*, pages 451–456.

Sennrich, R., Haddow, B., and Birch, A. (2016).

Improving neural machine translation models with monolingual data.
In *Proc of the Ann Meeting of the Assoc for Comp Linguistics*, pages 86–96, Berlin, Germany.

Sitbon, L., Bellot, P., and Blache, P. (2010).

Lisibilité et recherche d'information : vers une meilleure accessibilité intégration de la lisibilité au calcul de la pertinence.

Smith, E. (1961).

Devereaux readability index.
The Journal of Educational Research, 54 :289–303.

Références IV

Wubben, S., van den Bosch, A., and Krahmer, E. (2012).
Sentence simplification by monolingual machine translation.
In *Annual Meeting of the Association for Computational Linguistics*, pages 1015–1024.

Zeng, Q. T., Kim, E., Crowell, J., and Tse, T. (2005).
A text corpora-based estimation of the familiarity of health terminology.
In *ISBMDA 2006*, pages 184–92.

Zhao, S., Wang, H., and Liu, T. (2010).
Leveraging multiple MT engines for paraphrase generation.
In *COLING*, pages 1326–1334.