Rated Lexicon for the Simplification of Medical Texts

Anaïs Koptient

CNRS, Univ. Lille, UMR 8163 STL
Savoirs Textes Langage
F-59000 Lille, France
Email: anais.koptient.etu@univ-lille.fr

Natalia Grabar

CNRS, Univ. Lille, UMR 8163 STL
Savoirs Textes Langage
F-59000 Lille, France
Email: natalia.grabar@univ-lille.fr

***Abstract*—The purpose of our work is to create a rated lexicon in French useful for automatic text simplification of medical texts. Currently, the lexicon contains 11,272 pairs {*technical term*; *paraphrase*} for 6,937 different terms. This lexicon is built automatically using different methods. It is validated manually. Then, the lexicon is rated with several readability formulas and models in order to appraise the readability of terms and paraphrases. The lexicon will be exploited and tested within automatic simplification systems, and will be made available for the research community.**

Keywords–Medical and Health Text; Simplification; Lexicon; Paraphrases; Readability; Readability scores; Patients.

I. Introduction

The purpose of Automatic Text Simplification (ATS) is to make a given text more understandable for a group of persons, like children, people with pathologies, foreigners, people with no training in a specialized domain, etc. The last few years have seen a growing interest for the ATS, with the main bulk of work done in English and on general-language texts. Very little work exists on simplification of specialized texts, like medical texts, and on languages other than English. Nevertheless, these recent works helped the field to gain in maturity.

Several levels of simplification are distinguished:

- Lexical simplification, in which difficult words are replaced by the corresponding easier words. This kind of simplification is performed on the basis of lexical knowledge, such as synonyms, hyperonyms, definitions, etc. An example from the *SemEval 2012* challenge on English Lexical Simplification [1] is given in (1), in which the word *atrocities* is considered to be difficult to understand and is replaced by *cruelties*;

  (1) *Hitler committed terrible atrocities during the second World War.*
  Hitler committed terrible cruelties during the second World War.

- Syntactic simplification is done at the level of syntactic trees and has the purpose to reduce the syntactic complexity of sentences. Syntactically complex sentences can be transformed into simpler syntactic structures by using to deletion, insertion, separation, merging, and reordering. In example (2), borrowed from [2], the subordinate clause is separated from the main clause;

  (2) *While the law generally supports clampers operating on private land, Mr Agar claims CCSs sign was not prominent enough to be a proper warning.*
  The law generally supports clampers operating on private land. But Mr Agar claims CCSs sign was not prominent enough to be a proper warning.

- Semantic simplification implies that information can be reorganized or added to make the understanding easier thanks to the context [3]. Hence, the word *gabapentine* becomes easier to understand in (3);

  (3) *Gabapentine should be prescribed with caution to pregnant women.*
  Gabapentine medication should be prescribed with caution to pregnant women.

- Pragmatic simplification may imply that the structure of the text is modified [4], and its semantic cohesion becomes more global [5][6].

Currently, the researchers have identified different ways to simplify texts automatically: (1) approaches based on distributional probabilities, such as word embeddings [7][8], which permit to propose simpler candidates for a given word considered as difficult to understand; (2) approaches based on automatic translation systems [9][10], which consider the simplification as a monolingual translation task; (3) rule-based approaches [11][12], which design and exploit specifically defined simplification rules. Whatever the approach, the common point is the need for resources, such as dedicated simplification corpora, syntactic transformation rules, and lexical resources in which difficult words are associated with simpler synonyms, like {*atrocity*; *cruelty*}. Yet, synonyms are not the only lexical information necessary for the simplification. Hence, the few existing works on the typology of simplification [6][13] show that lexical substitution can also be performed with extended forms of abbreviations, hyperonyms, hyponyms, paraphrases and definitions. Such lexical resources must be reliable and propose simpler equivalents for technical terms.

The purpose of our work is to build such reliable lexical resource for French, with the focus on medical language and terminology. Therefore, we need (1) to identify lexical equivalents (synonyms, hyperonyms, definitions, etc.) for technical medical terms, and (2) to assign readability scores to technical terms and to their equivalents.

In our work, *term* or *technical term* correspond to terms that need to be simplified during the simplification process. They can correspond to syntactically simple (one word, like *comedo* or *hematuria*) or complex (more than one word, like *systemic lupus erythematosus*) sequences. *Paraphrases* or *equivalences* are the simplified layman versions with the same, or very close, meaning. Both elements are associated within the same pair {*technical term*; *paraphrase*}.

In what follows, we first present the methods designed for the identification of lexical equivalents for technical terms (Section II). We then describe the approaches for rating the lexicon (technical terms and their equivalents) according to the readability and evaluate the results (Section III). Finally, we conclude with some perspectives for future work (Section IV).

II. Identification of Lexical Equivalents for Technical Medical Terms

In this section, we introduce the corpora used and explain the methods proposed for the identification of lexical equivalents (synonyms, hyperonyms, definitions, etc.) for technical terms.

A. Corpora

We use the CLEAR corpus [14], which contains comparable documents differenciated by their technicality and difficulty. In this corpus, technical documents are associated with their simpler or simplified versions. The corpus contains 16,313 pairs of texts (over 57M word occurrences in technical texts and over 35M word occurrences in simplified texts), which are provided from three sources:

- *Drug leaflets* from the French ministry of health [15]. The technical part contains drug leaflets created for medical doctors, while the simple part contains patient package inserts that can be found in drug boxes. These two kinds of documents are created by pharmaceutical companies almost independently from one another;
- *Abstracts of systematic reviews* from the Cochrane collaboration [16]. The technical part contains technical abstracts, while the simple part contains the manually simplified versions of these technical abstracts;
- *Encyclopedia articles* from collaborative online encyclopedias. The technical part contains medicine-related articles from French Wikipedia [17], while the simple part contains the corresponding articles from the French children encyclopedia called Vikidia [18].

We also use a *forum* corpus collected from *masante.net*. This forum provides the possibility for users to ask health-related questions, which are answered by medical doctors. We exploit 6,139 answers available totaling 315,362 word occurrences.

B. Methods for Identification of Lexical Equivalents

We propose several methods for the identification of lexical equivalents (synonyms, hyperonyms, definitions, etc.) for technical terms. We also evaluate the extracted equivalents with the precision measure (percentage of correct equivalents among the extractions proposed by a given method). Each pair {*technical terms*; *paraphrase*} was validated manually by one person with training in NLP (Natural Language Processing) but no training in medicine. Table I summarises the extraction results provided by each method and their precision. In the second part of Table I, we indicate some existing work on acquisition of equivalents for medical terms in English [19][20] and in French [21][22]. The most known resource is the *Consumer Health Vocabulary* (CHV) in English [19], while there is no comparable resources in other languages, such as French. The methods exploited for the creation of CHV are both manual and automatic. Overall, CHV contains 141,213 unique layman terms, among which 11,641 terms are lexically different from their technical terms. This lexicon is the closest work to what we present in this section. Our lexicon currently contains 11,272 pairs {*technical terms*; *paraphrase*} for 6,937 different terms. Because of their specific linguistic function, abbreviations are not included in the lexicon. Besides, several other extractions need yet to be validated manually.

TABLE I. SUMMARY OF DIFFERENT METHODS PROPOSED: NUMBER OF CORRECT EXTRACTIONS AND THEIR PRECISION, AND THEIR COMPARISON WITH THE EXISTING WORK

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 [19]	11,641	–
English medical abbreviations [20]	785	95
French medical terms [21]	147	67
French medical terms [22]	109	66

1) Extraction of Equivalents from Parallel Aligned Sentences: Manually aligned parallel sentences from the CLEAR corpus are first manually annotated for transformations observed during the simplification of technical sentences. The annotation is done within the YAWAT annotator [23]. The annotations focus on several types of transformations, among which the most frequent are: (1) synonymy ({*excipients*; *composants*} *(*{excipients*;* components}*)*, {*céphalées*; *maux de têtes*} *(*{cephalalgia*;* headaches}*)*), (2) hyperonymy ({*clyndamicine*; *ce médicament*} *(*{clyndamicin*;* this drug}*)*), (3) hyponymy ({*benzodiazépines*; *bromazépam*} *(*{benzodiazepines*;* bromazépam}*)*), (4) part-of-speech shift ({*peuvent se manifester*; *apparition*} *(*{can appear*;* occurrence}*)*), (5) formal shift ({*des médicaments*; *un médicament*} *(*{drugs*;* drug}*)*). Once the transformations are annotated, we extract the equivalents which correspond to synonyms and hyperonyms, and which are the easiest to exploit during the simplification. This resource includes 626 technical terms with their equivalents. Due to the method, fully relying on manual annotation, this set of equivalents shows 100% precision.

2) Definitions of Technical Terms: Definition context of terms, like *est un (is a)* or *défini comme (defined as)* are exploited to extract definitions of medical terms. An example is given in (4). Technical terms are first detected and, if they occur within definition contexts, the entire sentence is extracted. 2,037 candidate definitions are extracted. After the validation, we keep 1,028 definitions (68% precision).

(4) *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 technique for blood vessels which are*

not visible with standard imaging.)

3) Reformulations of Technical Terms: Reformulations usually indicate that there are technical terms and that they are explained by the speaker [24]. We exploit several linguistic markers: (1) brackets like in (5), in which the technical word *hématurie (hematuria)* is reformulated in *trop de globules rouges dans vos urines (too much of red blood cells in urine)*; (2) explicit reformulation markers like *c'est-à-dire (that is (to say))*, *autrement dit (in other words)*, *l'équivalent (the equivalent)* or *encore appelé (also called)*. In example (6), the technical term *périménopause (perimenopause)* is reformulated in *période qui entoure la ménopause (period which surrounds the menopause)*.

(5) *Vous avez effectivement une hématurie (trop de globules rouges dans vos urines). (Indeed, you have hematuria (too many red blood cells in urine).)*

(6) *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.)*

This method provides 7,959 correct pairs {*technical term*; *paraphrase*} which overall precision is 60%. With this kind of method, it is also necessary to verify the direction of the relation: where is the technical term and where is its layman paraphrase. During the extraction, we consider that the longer sequence is the paraphrase, contrary to the technical term, which is usually a single-word expression or a noun phrase. This feature is also checked in Section III.

4) Word Morphology: In the biomedical language, word morphology may be indicative of technical terms and of their possible paraphrases, like in *myalgia*, composed of *myo- (muscle)* and *algia (pain)*, and meaning *muscle ache*. We exploit information on word morphology in two ways:

- The terms are first analyzed morphologically with Dérif [25] in order to transform them into morphological bases and affixes: *myocardique (myocardial)* is analyzed into *myo (muscle)* and *carde (heart)*. Then, we look into the corpus and search for syntactic groups that contain these words (*muscle* and *heart* in this example). In this way, we can find the sequence *heart muscle* meaning *muscle du coeur* in French. This method provides 1,128 paraphrases for technical terms with 86% precision;
- We start with a set of Latin and Greek affixes (430 prefixes and 103 suffixes) and their semantics, like *dipsy* meaning *thirst*, *a* meaning *absence/without*, *logy* meaning *study of*, or *angio* meaning *blood vessel*. We then combine every prefix with every suffix [26] to coin possible medical terms. In this way, we obtain 15,405 possible medical terms, which are then validated manually: this results in 1,939 terms (13% precision). Supposing that medical terms are compositional, we also combine the meaning of their morphological components for the creation of paraphrases: *angiologie (angiology)* is paraphrased in *étude des vaisseaux sanguins (study of blood vessels)*, while *adipsie (adipsy)* is paraphrased in *absence de soif (absence of thirst)*.

5) Expansion of Abbreviations: Abbreviations are commonly used in the medical language, like *LCR (CSF)* meaning *liquide cérébro-spinal (cerebrospinal fluid)* in example (7). Unless already known, abbreviations are difficult to understand by patients: it is then necessary to provide expanded forms of abbreviations. We extract the expanded forms of abbreviations with an adapted version of published algorithm [20], which processes two kinds of structures: *expanded form (abbreviation)* like in (7), and *abbreviation (expanded form)*. We extract 8,148 abbreviations with precision 94%.

(7) *On l'appelle aussi liquide cérébro-spinal (LCR). (It is also called cerebrospinal fluid (CSF).)*

6) Exploitation of an Online Medical Dictionary: We also exploit already available lexicons from online sources [27]. For each medical term, we keep the first sentence of the definition, which is expected to describe precisely the term. We obtain 1,165 additional medical terms and their paraphrases.

III. Computing the Readability of Technical Terms and of Their Equivalents

In this section, we compute the readability scores for technical terms and their layman equivalences. The purpose is (1) to assign the readability scores to each term and paraphrase, (2) to verify if paraphrases are indeed easier than technical terms, (3) if necessary, to switch the place of terms with their equivalents, which can be relevant with some automatic methods like reformulation extraction, and more specifically (4) to provide indication on simplicity of terms and their equivalents, which can be later used by simplification systems. For instance, some technical terms have more than one equivalent, which differ by their readability. In this situation, it is necessary to choose the equivalent which suits the best the simplification task.

Over one hundred readability formulas have been proposed by researchers [28], from which we choose just few for our work. These are linear regression formulas. They are mainly dedicated for rating the readability at the level of texts. These readability indexes are not considered to be very reliable and are often criticized [19]. Nevertheless, we consider that they can provide useful information on readability of terms and paraphrases. In the rest of this section, we first present the selected readability formulas and how we adapt them for the processing of terms and paraphrases (Section III-A). We then propose our computational readability models adapted to paraphrases, and based on a set of features and machine learning algorithms (Section III-B). The models are evaluated with Precision (correctness), Recall (exhaustiveness) and F-measure (harmonic mean of Precision and Recall). Finally, we present the results obtained with the readability models and indexes (Section III-C).

A. Linear Regression Readability Formulas

Dale index [29] is one of the first readability formulas proposed: $Dale = 0.15x1 + 0.04x2$, where *x1* represents the percentage of words missing from the basic vocabulary, and *x2* represents the average number of words per sentence. The higher *Dale* index, the less the text is readable. We adapt this formula to terms in French as follows: *x1* is the percentage of words missing form the Catach list [30], which is the French set with 400 basic words; and *x2* is the number of words in a given paraphrase or term. When applied to paraphrases, this formula provides readability scores between 0.08 (*être malade(being sick)*) and 15.4 (*éruption faciale, douleur articulaire, anomalies musculaires, fièvre (facial rash, articular pain, muscle abnormality, fever)*). The scores of terms are lower.

Kandel index [31] is the French adaptation of the very popular Flesch formula [32]: $Kandel = 207-(1.015*ASL)-(73.6*ASW)$, where *ASL* is the average number of words in each sentence, and *ASW* the average number of syllables. The index values are expected to fall between 0 and 100: 0 to 30 for texts difficult to understand, and starting from 70 for texts easily understandable by adults. In our experiments, we consider that *ASL* is the number of words in the paraphrase, and *ASW* the average number of syllables per word. When applied to paraphrases, the index scores are uneven and fall outside the expected scale, going from -188.58 (*hypertension intracrânienne bénigne (benign intracranial hypertension)*) up to 204.96 (*condylomes acuminés (acuminated condyloma)*).

Mesnager index [33] is a variant of the *Dale* index: $Mesnager = (1/2 * AC) + (1/3 * P)$, where *AC* is the percentage of words missing from the basic vocabulary [30], and *P* the average number of words in sentences. The index values are supposed to be between 6 (easy text) and 25 (difficult text). In our case, we consider that *P* is the number of words in paraphrases and terms. When applied to our data, the formula provides scores between 0.66 (*être malade (being sick)* or *point noir (blackhead)*) and 69.3 (*éruption faciale (facial rash)*, *douleur articulaire (articular pain)*, *anomalies musculaires (muscle abnormality)*, and *fièvre (fever)*).

Sitbon index [34] is one of the rare formulas designed for sentences (and not for texts): $Sitbon = 1.12 * ADV - 0.69 * CON + 6.48 * cohesion + 15.58$, where *ADV* and *CON* are, respectively, the number of adverbs and conjuctions, and *cohesion* is the number of phonemes divided by the number of letters. There is no reference scale of values for the *Sitbon* index. When applied to our data, the index provides scores between 18.05 (*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)*) and 25.37 (*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)*). We can see that the scale of values is very narrow and offers reduced discrimination of readability.

Smith index [35] is also adapted to sentences: $L = -6.49 + 1.56WL + 0.19SL$, where *WL* is the average number of letters in words, and *SL* is the number of words in the sentence. When applied to our data, the formula shows scores between -1.44 (*étude de l'os (study of bone)*) and 17.29 (*concrétions gastro-intestinales (gastrointestinal concretions)*). Contrary to other indexes, difficult paraphrases are not the longer ones but rather those composed of polylexical units, like *gastro-intestinal (gastrointestinal)*.

B. Computational Readability Models

For designing the computational readability models, we choose the descriptors mainly issued from the existing typology [36]. The purpose is to design a set of descriptors easy to compute and to use:

- *number of letters*, usually indicating the length, and complexity, of terms and of their equivalents;
- *number of phonemes*. To obtain the number of phonemes, we use the database Lexique3 [37]. It provides over 140,000 French lemmas and associated information, such as their phonetic transcription, number of syllables, and part-of-speech tag. For words missing in Lexique3, we use the Epitran module [38] adapted to French;
- *number of syllables*. Lexique3 is also used to obtain the number of syllables. For words missing in Lexique3, we use Epitran and then their syllabation [39];
- *cohesion between phonemes and spelling* corresponds to the ratio between the number of phonemes and number of letters. It provides values between 0 and 2: 0 if no difference, 1 if one or two differences, and 2 if more than two differences. Words with higher values of cohesion are supposed to be less readable;
- *frequency* is also obtained from Lexique3. For words missing in Lexique3, we fix the frequency to 0 because these are supposed to be rare words;
- *presence in the Catach list [30]*, which is the basic set of French words;
- *syllable components*, which corresponds to three complexity levels according to the structure of syllables (coined with consonants C, vowels V and semi-consonants Y) and their frequency. For instance, syllables like *CYV, V, CVC, CV* are very frequent in French, while syllables like *CCVC, VCC,VC, YV, CVY* are much less frequent in French.

We have to predict two classes for terms and equivalents: *simple* and *difficult*. Training of the biclass models is done on independent reference data: manually rated medical lexicon annotated according to the difficulty of words [40]. This lexicon contains 29,641 medical words. Three classifiers (MultiLayer Perceptron MLP, Decision Tree DT and Random Forest RF), implemented within the Python library ScikitLearn [41], are used. Table II indicates Precision, Recall and F-measure obtained during the training with a 10-fold cross-validation set. We can see that all classifiers show good results, MLP being the best in this task with overall results over 90%.

TABLE II. RESULTS OF THE READABILITY MODEL ON TRAINING REFERENCE DATA WITH 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

The models are next applied to terms and their paraphrases from the lexicon. The more the prediction is close to 0 the more difficult is the sequence, and the more it is close to 1 the simpler is the sequence. When the sequence contains

TABLE III. EXAMPLES OF RATING OF TECHNICAL TERMS AND OF THEIR EQUIVALENTS

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

more than one word, which is the majority of cases, models are first applied to each non-grammatical word, and then we compute the average probability of the whole sequence to be classified as simple or complex. The probabilities of three algorithms are taken into account individually. For instance, in *abaissement de la température (decrease in temperature)*, all the algorithms predict that *abaissement (decrease)* is simple (with probability value 1) and that *température (temperature)* is simple (with probability 1). This gives the average score 1 for each algorithm, and the term is considered as simple by all of them. As for *ablation de l'abdomen (ablation of abdomen)*, MLP and RF predict that the two words of the paraphrase are simple (probability 1), while DT predicts that *ablation* is simple and *abdomen* is difficult. This gives the average score 1 for MLP and RF, and 0.5 for DT. Overall, this term is also considered as simple but with lesser probability.

C. Results

The result of this step is that technical terms and paraphrases are rated for their readability with the five classical readability indexes (*Dale, Kandel, Mesnager, Sitbon* and *Smith*) and by the proposed computational readability models. In Table III, we present some examples of technical terms and of their equivalents, and indicate their readability scores. In the first line, we indicate the interpretation of the readability values according to indexes and models. For instance, with *Dale*, high scores are expected to be associated with difficult terms, while low scores are expected to be associated with simple terms. The *Sitbon* index is rather sensitive to long terms and paraphrases. In the examples provided, technical terms precede the paraphrases. For instance, *comédon (comedo)* is recognized to be difficult to undestand by all measures: *Dale*, *Mesnager*, *Sitbon* and *Smith* indexes are high, *Kandel* is low, and the three computational models *MLP, DR, RF* show the value 0. As expected, its paraphrase *point noir (blackhead)* is recognized to be easy to understand: *Dale*, *Mesnager*, *Sitbon* and *Smith* indexes are low, *Kandel* is high, while the three comutational models *MLP, DR, RF* show the value 1. The picture may be different with other pairs $\{$*term*; *paraphrase*$\}$. For instance, in the pair $\{$*vomissement (vomiting)*; *être malade (being sick)*$\}$, both elements are considered as understandable by computational models and *Sitbon*, while other indexes consider that the paraphrase *être malade (being sick)* is simpler than the term *vomissement (vomiting)*. The pairs, in which terms are paraphrased with long sequences, may be more difficult to be rated by the indexes and models. This is the case of *système endocrinien (endocrine system)* and its paraphrase *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)*. Hence, the length of the paraphrase may introduce additional readability factor, which should also be considered in chosing the paraphrases for simplification. Overall, indexes and models provide useful information for the selection of lexical substitutes for technical terms.

The scores also permit us to compare the readability within pairs and indicate that order of terms and their paraphrases is correct. In few cases, the length of paraphrases decreases their readability, but overall their readability remains acceptable.

IV. Conclusion

Automatic text simplification is an NLP field whose purpose is to make texts more easily understandable by common readers. While an important progress has been done in this field, the main barrier is still related to the availability of suitable data, such as corpora and lexica. We propose a set of experiments designed for the creation of a lexicon with French technical medical terms and their layman paraphrases. Several approaches and methods are developed and applied for the automatic extraction of paraphrases. The results from each method are evaluated with precision metric and usually show that the extractions are reliable with over 68% precision. Overall, the lexicon contains 11,272 pairs $\{$*technical term*;

paraphrase} for 6,937 different technical terms. Terms and paraphrases from this lexicon are then rated for their readability with several adapted readability indexes and with specifically designed computational models. Globally, we observe that paraphrases are indeed easier to understand than technical terms. This rated lexicon will be exploited by simplification systems and we expect that readability scores will help to choose the best lexical substitutions. The lexicon will be made available for the research community.

Acknowledgment

This work was funded by the French National Agency for Research (ANR) as part of the CLEAR project (Communication, Literacy, Education, Accessibility, Readability), ANR-17-CE19-0016-01.

References

[1] L. Specia, S. Jauhar, and R. Mihalcea, "Semeval-2012 task 1: English lexical simplification," in **SEM 2012*, 2012, pp. 347–355.

[2] A. Siddharthan, "Syntactic simplification and text cohesion," *Research on Language & Computation*, vol. 4, no. 1, pp. 77–109, 2006.

[3] L. Brouwers, D. Bernhard, A.-L. Ligozat, and T. François, "Syntactic sentence simplification for French," in *PITR workshop*, 2014, pp. 47–56.

[4] C. Vettori and O. Mich, "Supporting deaf children's reading skills: the many challenges of text simplification," in *ASSETS*, 2011, pp. 1–2.

[5] H. M. Caseli, T. F. Pereira, L. Specia, T. A. S. Pardo, C. Gasperin, and S. M. Aluisio, "Building a Brazilian Portuguese parallel corpus of original and simplified texts," in *CICLING*, 2009, pp. 1–12.

[6] D. Brunato, F. Dell'Orletta, G. Venturi, and S. Montemagni, "Defining an annotation scheme with a view to automatic text simplification," *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 and of the Fourth International Workshop EVALITA 2014*, pp. 87–92, 2014.

[7] G. Glavas and S. Stajner, "Simplifying lexical simplification: Do we need simplified corpora?" in *ACL-COLING*, 2015, pp. 63–68.

[8] Y.-S. Kim, J. Hullman, M. Burgess, and E. Adar, "SimpleScience: Lexical simplification of scientific terminology," in *EMNLP*, 2016, pp. 1–6.

[9] S. Zhao, H. Wang, and T. Liu, "Leveraging multiple MT engines for paraphrase generation," in *COLING*, 2010, pp. 1326–1334.

[10] S. Nisioi, S. Stajner, S. P. Ponzetto, and L. P. Dinu, "Exploring neural text simplification models," in *Ann Meeting of the Assoc for Comp Linguistics*, 2017, pp. 85–91.

[11] J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait, "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, Jun. 1999. [Online]. Available: https://www.aclweb.org/anthology/E99-1042

[12] J. De Belder, K. Deschacht, and M.-F. Moens, "Lexical simplification," in *ITEC*, 2010, pp. 1–4.

[13] A. Koptient, R. Cardon, and N. Grabar, "Simplification-induced transformations: typology and some characteristics," in *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 309–318. [Online]. Available: https://www.aclweb.org/anthology/W19-5033

[14] N. Grabar and R. Cardon, "Clear – simple corpus for medical French," in *Workshop on Automatic Text Adaption (ATA)*, 2018, pp. 1–11.

[15] Base de données publique des médicaments (*Public Database of Drugs*). Ministère des Solidarités et de la Santé [Health and Solidarities Ministry]. [Online]. Available: http://base-donnees-publique.medicaments.gouv.fr/

[16] Cochrane reviews. Cochrane Library. [Online]. Available: http://www.cochranelibrary.com/

[17] Wikipédia l'encyclopédie libre (*Wikipedia the free encyclopedia*). Wikipédia. [Online]. Available: https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal

[18] Vikidia. Vikidia. [Online]. Available: https://fr.vikidia.org/wiki/Vikidia:Accueil

[19] Q. T. Zeng, E. Kim, J. Crowell, and T. Tse, "A text corpora-based estimation of the familiarity of health terminology," in *ISBMDA 2006*, 2005, pp. 184–92.

[20] A. S. Schwartz and M. A. Hearst, "A simple algorithm for identifying abbreviation definitions in biomedical text," in *Pacific Symposium on Biocomputing*, 2003, pp. 451–456.

[21] L. Deléger and P. Zweigenbaum, "Paraphrase acquisition from comparable medical corpora of specialized and lay texts," in *Ann Symp Am Med Inform Assoc (AMIA)*, 2008, pp. 146–50.

[22] B. Cartoni and L. Deléger, "Découverte de patrons paraphrastiques en corpus comparable: une approche basée sur les n-grammes (*Discovery of paraphrastic patterns in comparable corpora: an n-gram based approach*)," in *Traitement Automatique des Langues Naturelles (TALN)*, 2011.

[23] U. Germann, "Yawat: Yet another word alignment tool," in *ACL*, 2008.

[24] E. Antoine and N. Grabar, "Exploitation de reformulations pour l'acquisition d'un vocabulaire expert/non expert (*Exploitation of reformulations for the acquisition of expert/non-expert vocabulary*)," in *Traitement Automatique des Langues Naturelles (TALN)*, 2016.

[25] F. Namer, *Morphologie, Lexique et TAL : l'analyseur DériF (Dérif Analyzer). TIC et Sciences cognitives*. London: Hermes Sciences Publishing, 2009.

[26] N. Kloehn *et al.*, "Improving consumer understanding of medical text: Development and validation of a new subsimplify algorithm to automatically generate term explanations in english and spanish," *J Med Internet Res*, vol. 20, no. 8, p. e10779, Aug 2018. [Online]. Available: http://www.jmir.org/2018/8/e10779/

[27] Fondation contre le cancer (*Fundation against Cancer*). [Online]. Available: https://www.cancer.be/lexique

[28] T. François, "Les apports du traitement automatique du langage à la lisibilité du français langue étrangère (*Contributions of Natural Language Processing to the readability of French as a foreign language)*," PhD thesis, Université Catholique de Louvain, Louvain, 2011.

[29] E. Dale and J. S. Chall, "A formula for predicting readability," *The Journal of Educational Research*, vol. 27, no. 11-20, pp. 37–54, 1948.

[30] N. Catach, F. Jejcic, and E. H. C. N. de la Recherche Scientifique), *Les listes orthographiques de base du franais (LOB) : les mots les plus frquents et leurs formes flchies les plus frquentes (Basic lists of spelling in French: the more frequent words and their more frequent inflected forms)*. Paris, France: Nathan, 1984.

[31] L. Kandel and A. Moles, "Application de l'indice de flesch la langue francaise [applying flesch index to french language]," *The Journal of Educational Research*, vol. 21, pp. 283–287, 1958.

[32] R. Flesch, "A new readability yardstick," *Journ Appl Psychol*, vol. 23, pp. 221–233, 1948.

[33] J. Mesnager, "Lisibilité des textes pour enfants : un nouvel outil ? (*Readability of texts for children: a new tool?*)," 1989.

[34] L. Sitbon, P. Bellot, and P. Blache, "Lisibilité et recherche d'information : vers une meilleure accessibilité (*Readability and Information Seeking: towards a better accessibility*)," 10 2010.

[35] E. Smith, "Devereaux readability index," *The Journal of Educational Research*, vol. 54, pp. 289–303, 1961.

[36] N. Gala, T. François, D. Bernhard, and C. Fairon, "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*, Marseille, France, 2014, pp. 91–102. [Online]. Available: https://www.aclweb.org/anthology/F14-1009

[37] B. New, C. Pallier, L. Ferrand, and R. Matos, "Une base de données lexicales du francais contemporain sur internet : Lexique (*A lexical database for contemporary french: LEXIQUE*)," *Annee Psychologique - ANNEE PSYCHOL*, vol. 101, pp. 447–462, 01 2001.

[38] D. R. Mortensen, S. Dalmia, and P. Littell, "Epitran: Precision G2P for many languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. C. C. chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds. Paris, France: European Language Resources Association (ELRA), May 2018.

[39] C. Pallier, “Syllabation des représentations phonétiques de brulex et de lexique (*Syllabation of phonetics representation of Brulex and Lexique*),” 1999.

[40] N. Grabar and T. Hamon, “A large rated lexicon with french medical words,” in *LREC*, 2016.

[41] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.