Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons

Núria Gala[1], Thomas François[2], Cédrick Fairon[2]

(1) LIF-CNRS, Aix Marseille Université, 163 av. de Luminy case 901,
13288 Marseille Cedex 9, France
(2) CENTAL, Université Catholique de Louvain, Place Blaise Pascal 1,
1348 Louvain-la-Neuve, Belgique
E-mail: nuria.gala@lif.univ-mrs.fr, {thomas.francois}{cedrick.fairon}@uclouvain.be

Abstract

The readability of a text depends on a number of linguistic factors, among which its lexical complexity. In this paper, we specifically explore this issue: our aim is to characterize the criteria that make a word easy to understand independently of the context in which it appears. Yet such a concern must be addressed in the context of particular groups of individuals. In our case, we have focused on language production from patients with language disorders. The results obtained from corpus analysis enable us to define a number of variables which are compared to information from existing resources. Such measures are used in a classification model to predict the degree of difficulty of words and to build a lexical resource, called *ReSyf*, in which the words and their synonyms are classified according to three levels of complexity.

Keywords: lexical resource, readability, simplification, natural language processing, language model.

1. Introduction

There has been a significant number of works on the readability and simplification of texts over the last 80 years. Most of them take into account the lexicon in an assessment of text difficulty. For instance, Flesch (1948) used the number of syllables per word as a measure of word complexity. Smith (1961) instead suggested using the mean number of letters, since this is easier for a computer to calculate. Stenner and Burdick (1997) predicted text difficulty from the logarithm of word frequencies.

However, although all these studies were concerned with the impact of the lexicon on text difficulty, they did not directly assess the complexity of the lexicon. Efforts at this level were more concerned with designing lists of 'easy' words. Such lists have been produced for teaching purposes in different languages, relative to a first language (L1) or a second language (L2). Among them, some of the most well-known are, for English, the *Teachers' Book of Words* (Thorndike, 1921) and the *Basic English* (Ogden, 1930) and, for French, *Le Français Fondamental* (Gougenheim, 1958) and the *Listes Orthographiques de Base du Français* (Catach, 1985).

Although these lists were subsequently used for text readability purposes (Dale and Chall, 1948), their use presents several limitations in terms of assessing the difficulty of a whole lexicon. First, the lists are based on a single criterion, such as the

frequency of words (Thorndike, 1921), or the percentage of words known by 80% of schoolchildren from the fourth grade (Dale, 1931). More importantly, their coverage is generally limited to a set of a few hundred 'easy' words, making them too restricted to be used, for instance, in text simplification systems. The problem of coverage is accentuated as the vocabulary of a language is in constant evolution.

Therefore, it appears that a more integrated approach, using Natural Language Processing (NLP) techniques, could be suitable for automatically predicting the difficulty of words.

To our knowledge, the only readability study that proposes a formula directly at the lexicon level is that of Bormuth (1966). He first used the cloze test procedure[1] to yield a corpus of 20 educational texts annotated in terms of difficulty at the word level. Then, he modelled word difficulty with four variables: the number of syllables, the number of letters, a frequency index, and the word depth as defined by Yngve (1962). When combined, these four variables produced a multiple correlation coefficient (R) of 0.505, a far lower score than that obtained by the text level model (R = 0.934). From this study, it appears that predicting the difficulty of words is surprisingly harder than predicting text difficulty.

In this paper, we first explore a larger set of variables to predict the degree of difficulty of a word. Then, using these scores, we build a synonym lexicon where each word has a difficulty index. Such a resource is to be used (1) by humans for language comprehension or production and (2) by a language model for automatic simplification. To our knowledge, no existing lexical resource, except for graded scholar word lists, offers its users the possibility to select words according to their degree of difficulty.

The article is organized as follows. In the next section, we discuss which characteristics of a word make it simple or difficult, according to psycholinguistic studies and linguistic variables that we have defined. In section 3, we describe the resources we use to compare our features on two sets of words: (a) words used in a given task by patients affected by Parkinson's disease, (b) words from a large general lexical list. In Section 4, we report experiments and methodology to design a first gold-standard graded list and a model of lexicon difficulty. Finally, we conclude with some remarks on the limitations of our present approach and proposals for future work.

2. How simple can a word be?

Identifying how simple a word can be has been of interest to psycholinguists for many

[1] This test, designed by Taylor (1953) to measure reading comprehension, requires readers to read a text with regular blanks (one every five words) and fill in as many blanks as possible.

years. Experiences of the complexity of words with regards to various recognition tasks (lexical decision, semantic categorization, etc.) have been intensely reported in the literature (Ferrand, 2007). One of the main findings is the word frequency effect: a high-frequency word is recognized more easily than one of low frequency. The close correlation between frequency and difficulty has been highlighted in many studies (Howes and Salomon, 1951; Brysbaert et al., 2000).

Other word-level effects have been stressed in psycholinguistic literature, such as the familiarity effect (Gernsbacher, 1984), the age of acquisition effect (Morrison and Ellis, 1995), the orthographic neighbour effect (Andrews, 1997), the length of words (O'Regan and Jacobs, 1992), etc. Most of these effects are indeed correlated with the difficulty of texts (François and Fairon, 2012) and are likely to be also a valuable source of information for a model of word complexity.

A second source of information about word simplicity comes from linguistic studies on levels lower than the word unit: morphemes or phonemes. Intra-lexical factors, such as familiarity of phonemes, regularity in pronunciation, fixed stress, consistency of the sound-script relationship, inflexional and derivational regularity, morphological transparency, generality, register neutrality, or number of meanings per form, affect vocabulary learning (Laufer, 1997). For Schreuder and Baayen (1997), the number of morphemes correlated with the size of the derivational family has an impact on visual word recognition.

To various extents, all these factors combine to explain word difficulty. It is acknowledged that the combination is dependent on a given group (or 'class') of individuals (François, 2012). What may be simple for one group may not be for another, especially since there is a wide variety of readers who do not have the same needs. However, we believe that, in order to describe how simple words can be, there are some general characteristics that can be related to fine-grained linguistic criteria. NLP methods are useful in formalizing such features and checking them on large amounts of data.

For the purposes of this study, we have identified a set of variables from the two following families:

- *Intra-lexical variables*: (1) number of letters, (2) number of phonemes, (3) number of syllables, (4) syllable structure, (5) consistency of sound-script relationship, (6) spelling patterns, (7) number of morphemes, (8) composition, and (9) affix frequency (for derived word).

- *Psycholinguistic variables*: (10) phonological neighbourhood, (11) orthographic neighbourhood, (12) abstract-concrete or imageability, (13) lexical frequency, (14) size of the derivational family, (15) absence/presence from Gougenheim list (Gougenheim et al., 1964), etc.

To check how these variables relate to difficulty, we performed two experiments. First, we computed their values on a *simplified* language corpus (see Section 4.2 for implementation details of the variables) and compared these results with values obtained from a general language lexical database. In seeking a corpus attesting some simplified language, we considered that linguistic productions from people with speech-related disorders might be a good start for observing 'simple' vocabulary. We therefore collected a corpus containing language productions of sufferers of Parkinson's disease (other types of speech-related disorders might be considered in the future). Second, we analysed how those variables vary within a lexicon of graded words for French, intended for schoolchildren (see Section 4.2.3).

3. Resources

In this section, we present the four resources used in our experiments. First, we describe a corpus with simple language productions. Second, we introduce a lexical database for French, Lexique3 (New et al., 2005), that is a representation of the general vocabulary. These two resources enable us to test some variables that potentially account for simple words. We also describe Manulex (Lété et al., 2004), a list of word frequencies at various school grade levels. Lastly, we present JeuxDeMots (Lafourcade, 2007), a lexical network that helped us to build *ReSyf*, our list of graded synonyms.

3.1 Parkinson corpora

The general public mainly recognizes Parkinson's disease through its motor symptoms (rest tremor, akinesia, and rigidity). However, the pathology may also entail language and speech impairments[2], namely dysarthria (Pinto et al., 2010), which includes hypophonia (reduced voice volume), monotone speech, and difficulties with articulation of certain sounds and syllables, as well as increased frequency and duration of hesitations and pauses (McNamara, 2010). Sentence structures are simplified (shorter), with an increase in the ratio of open-class items (nouns, verbs, adjectives, and adverbs) to close-class items (determiners, prepositions, conjunctions, etc.).

For our study, we used a corpus of twenty recordings from twenty Parkinson's patients describing the same picture (a short scene of an everyday situation)[3]. Patients were recorded whilst in 'off state', that is, with no medication that could have alleviated the effects of the disease.

After transcribing the twenty recordings, we obtained a corpus of 2,271 tokens that

[2] http://www.sciencedaily.com/releases/2011/02/11020262.htm

[3] The authors are grateful to S. Pinto from the Laboratoire Parole et Langage (LPL-CNRS, Aix Marseille Université) for providing the corpora and valuable insights on the disease.

we tagged using TreeTagger (Schmid, 1994). All marks of disfluencies, except repeated words, were removed (hesitations, truncated words, etc.). The average number of words per file was 113, the shortest file contained 42 words and the longest 233.

3.2 A lexical database for general French words

Lexique3[4] (New et al., 2005) is a free lexical database containing 142,728 words (47,342 correspond to a lemma; the other entries are inflected forms). Each word is described with phonological and morphological information (phonetic transcription, part of speech, morphological features [gender, number, tense, etc.], number of phonemes, number of syllables, syllable structure, number of morphemes, etc.). The database also provides estimates on the frequencies of occurrence of the words in books and film subtitles.

Figure 1 displays an example of the information available for the entry *armures* ('armours'):

ortho	phon	lemma	pos	gender
armures	aRmyR	armure	NOM	fem
number	V morpho	freq bks	freq films	**nb phon**
plu	-	5.46	8.11	**5**
struct lett	struct pho	syllables	**nb lett**	**nb syll**
VCCVCVC	VCCVC	aR-myR	**7**	**2**
sy struct pho	sy struct lett	nb homoph	nb homogr	**nb morph**
VC-CVC	ar-mu-re	1	0	**1**

Figure 1: The entry *armures* ('armours') from Lexique3.

Only some of the most significant fields are presented here, in the following order: spelling form, phonemic form, lemma, part-of-speech, gender, number, verbal morphology (tense, etc.), frequency estimated from a book corpora, frequency computed from film subtitles, number of phonemes, letter structure, phonemic structure, syllables, number of letters, number of syllables, syllable structure (phonemes) and syllable structure (letters), number of homophones, number of homographs, and number of morphemes.

3.3 A lexicon with scholar levels

To obtain a list of graded words, we used Manulex[5] (Lété et al., 2004), a list of French words whose frequencies have been extracted from primary school textbooks. For a

[4] http://www.lexique.org

[5] http://www.manulex.org

given word, the authors computed several measures (raw frequency, frequency of use over one million words, dispersion index and standard frequency index) for the three following levels of education:

- First year of primary school (children of 6 years old).
- Second year of primary school (7 years old).
- The three following years of primary school (8 to 10 years old).

Figure 2 provides an example of four entries, *pomme* ('apple'), *vieillard* ('old man'), *patriarche* ('patriarch') and *cambrioleur* ('burglar'). Only the raw frequency of each word per level of education is shown:

lemma	pos	Fq Lvl	Fq Lvl	Fq Lvl
pomme	N	724	306	224
vieillard	N	-	13	68
patriarche	N	-	-	1
cambrioleur	N	2	-	33

Figure 2: Sample of four entries in Manulex.

For the purpose of building a list of graded words, we transformed the frequency distributions over the three levels into a class system where a word can be assigned to only one class. As a first approach, we defined three classes corresponding to the three levels of education listed above and it was assumed that a given word would belong to the textbook level where it was first observed (e.g. level 1 for *pomme* and *cambrioleur*, but level 3 for *patriarche*). This straightforward classification has obvious shortcomings. For instance, it assigns the same level (level 1) to the words *pomme* and *cambrioleur* from Figure 2, whereas they present very different frequency distributions.

From this example, it seems that using a more complex function to transform the frequency distributions might produce a better classification. The idea is to give a different value to words, such as *pomme* – those that are more frequent at level 1 than at the other levels – and words such as *cambrioleur* that rather belong to levels 2 and 3. We thus experimented with the following formula:

$$N_c = N + e^{-r}, \quad \text{where} \quad r = \frac{\sum_{k=1}^{i} U_k}{\sum_{i+1}^{N} U_k}$$

N_c is a continuous score that is used at the word difficulty level instead of N, the level predicted by our first simple method describe above. N_c is obtained by summing N and a quantity e^{-r} that is inferior to 1 and is exponentially related to the ratio of the frequencies U_k at level k.

However, using this new scale did not lead to significant improvement for the

experiments described in Section 4.2, so we decided to use the simple approach throughout the paper. After applying the simple function and deleting grammatical words, we thus obtained a list containing 19,037 lemmas from Manulex, distributed as follows: 5863 words (31%) corresponding to level 1, 4023 words (21%) for level 2 and 9151 (48%) words for level 3.

At this stage, we compared the lemma list from the Parkinson corpora to the graded list obtained from Manulex, and the results were the following: 94.30% of the words in our corpora are tagged as belonging to the level 1 of Manulex, 1.45% are tagged as level 2, while only 1.63% belong in level 3 (the remaining 2.62% correspond to tagging errors, i.e. words tagged differently in the corpus and in Manulex). This confirms that the Parkinson list contains simple language productions.

3.4 A semantic network

JeuxDeMots[6] (JdM) is a freely available lexical network that is under development in the framework of a game for leveraging crowd-sourcing (Lafourcade, 2007). Given a trigger word, the game consists of proposing related words corresponding to a specific semantic or thematic relation. The resulting resource contains 163,543 words (in May 2013) with at least one lexical relationship (associated term, synonym, antonym, agent, patient, etc.).

Figures 3 and 4 display the information collected for the word *cambrioleur* ('burglar'). There are 114 thematic associations (*cheater*, *break in*, *thief*, *robbery*, *steal*, etc.) in which this word has been the trigger (Figure 3).

There are 71 relations (Figure 4) in which this word has been the target when asking for, line 4, 'agent of the verb *steal*', line 5 'who could hurt with a *weapon*', line 6 'synonym of *thief*', etc.

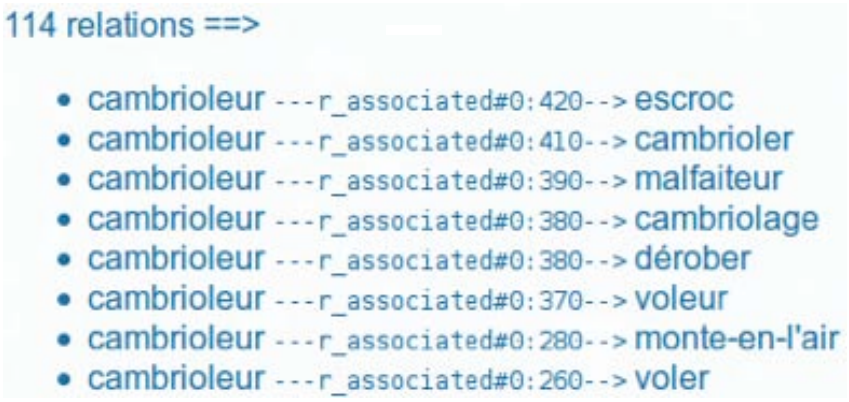

Figure 3: Outgoing relations in Jeux de Mots

[6] http://www.jeuxdemots.org/

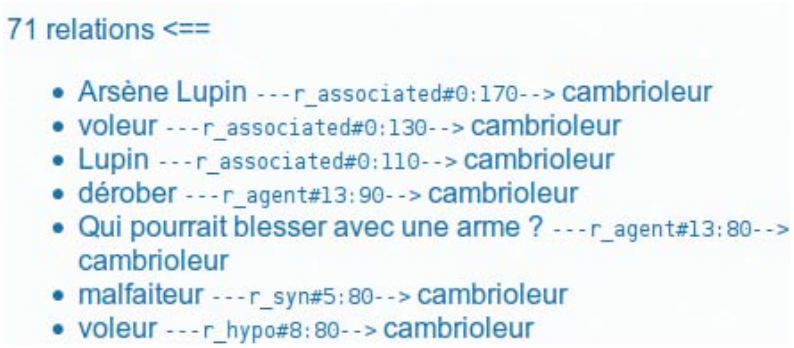

Figure 4: Ingoing relations in JeuxDeMots.

4. Building a graded synonym lexicon

Automatic acquisition of linguistic knowledge from corpora (raw texts or lexical resources) is a widespread trend in NLP. Over the last decades, many unsupervised and semi-supervised approaches have become a real alternative to manual development – too costly and time consuming (Gala and Lafourcade, 2011). More recently, collaborative approaches have emerged, based on the principle of sharing contributions (Calzolari, 2013), especially through games with a purpose (*gwap*), the lexical network JdM being an outstanding example of this trend.

For the purpose of creating a graded synonym lexicon, we first acquired information from Manulex in order to obtain a gold-standard list of graded words. Second, we implemented some of the identified intra-lexical and statistical features in order to automatically grade words outside our gold-standard list.

4.1 Acquiring information from existing resources to establish a gold-standard list of graded synonyms

We have indications that Manulex offers accurate difficulty classification: 94.3% of the words from the Parkinson's patients corpora correspond to level 1, which is consistent with what we know about language productions of Parkinson's patients. Therefore, we consider that Manulex grading can be used as a gold standard to create a first list of words with graded synonyms.

To this aim, we checked whether the 19,037 words of Manulex could be linked to synonyms in the lexical network JdM. From the initial 19,037 words in Manulex, 17,870 (93.87%) were present in JdM (the remaining words can be present in JdM, but with no known synonym relation). The distribution by level is as follows:

Level	Proportion	Counts
1	30.1%	5,375
2	21.0%	3,755
3	48.9%	8,740
1–3		17,870

Figure 5: Distribution of Manulex words in JdM.

From this list of 17,870 words, we gathered their synonyms in JdM: 10,975 have at least one synonym with a level in Manulex (we temporarily removed words absent from Manulex; they will be graded with our difficulty model). We obtained 12,687 graded synonyms, distributed as follows:

Level	Proportion	Counts
1	35.3%	4,477
2	21.7%	2,749
3	43,0%	5,461
1–3		12,687

Figure 6: Distribution of synonyms by level.

Figure 7 shows a sample of such a graded list containing synonyms from JdM along with their levels from Manulex:

Armure(1): protection(1), cuirasse(2), harnais(3)

Piétiner(2): marcher(1), fouler(3), piaffer(3), trépigner(3)

Patriarche(3): chef(1), père(1), vieillard(2)

Cambrioleur(1): malfaiteur(3), voleur(1), aigrefin(3)

Figure 7: Sample of ReSyf entries (*armour, protection, breastplate, harness; trample, walk, stamp one's feet, paw the ground; patriarch, chief, father, old man; burglar, criminal, thief, crook*).

We consider this list of graded words our gold-standard. Words absent from this list will be graded using our system for the automatic assessment of lexicon difficulty.

4.2 Towards a difficulty model for lexicon

This section presents the difficulty model we used to assess the difficulty of synonyms absent from Manulex. First, we detail which predictors of the lexicon complexity were implemented and how (at the time of writing this paper, only some had been tested). Then, we report two experiments performed on our three resource corpora that aimed to better understand which features are the most useful in predicting lexicon difficulty. Finally, we describe the model designed to assign a word to one of the Manulex levels.

4.2.1 Predictors of lexicon difficulty

As mentioned in Section 2, a large number of lexical predictors have been described in the literature. We implemented several of them, as follows:

a) Intra-lexical variables

(1) *number of letters*: we counted the number of alphabetical characters.

(3) *number of syllables*: we adopted a hybrid syllabification method. For words included in Lexique3, we used the gold syllabification included in the dictionary. For all other words, we generated API phonetic representations with espeak[7], and then applied the syllabification tool provided with Lexique3 (Pallier, 1999). The accuracy of this combined process exceeded 98% on a small test list.

(2) *number of phonemes* and (4) *syllable structure*: obtained from the syllabification system. For the syllable structure, we defined three categories of increasing difficulty, using their frequencies in the Parkinson corpus as a criterion: the most frequent structures (CYV, V, CVC, CV)[8], a group of less frequent structures (CCVC, VCC, VC, YV, CVY, CYVC, CVCC, CCV) and a final group containing only rare combinations.

(5) *consistency of sound-script relationship*: computed by comparing the number of letters and phonemes. We parameterized the output as three possible outcomes: 0 for complete transparency; 1 for a difference not higher than 2 characters, and 2 for words particularly obscure (difference higher than 2 characters).

(6) *spelling patterns*: defined as five categories of difficult patterns:

- double vowels (‘oo’, ‘éé’),

- double consonants (‘bb’, ‘cc’, ‘ff’, ‘gg’, ‘ll’, ‘mm’, ‘nn’, ‘pp’, ‘rr’, ‘ss’, ‘tt’),

- other digraphs in French (‘ck’ and ‘qu’ [k], ‘ch’ and ‘sh’ [ʃ], ‘ph’ [f], ‘gn’ [ɲ]),

- nasal vowels written with digraphs (‘an’ [ɑ̃], ‘in’ [ɛ̃], ‘on’ [õ], ‘un’ [oẽ])

- oral vowels written with digraphs (‘ai’ [e], ‘au’ [o], ‘eu’ [œ], ‘ou’ [u]).

There is work in progress concerning the remaining variables:

(7) *number of morphemes* and (8) *composition*: the hypothesis being that constructed words are more difficult to grasp.

(9) *affix frequency* on derived words: the difficulty of a derived word may depend on the frequency of the affix. In French, some affixes are very productive (*-age* with verbal basis as in *lavage* [‘wash’], *balayage* [‘weep’], tournage [‘filming’], etc.). Other affixes are quite rare (*-is* as in *treillis* [‘canvas’] or *tournis* [‘dizziness’]). The effect of affix frequency might have an impact on the level of difficulty of a word.

b) Psycholinguistic variables

(11) *orthographic neighbours*: computed from a list of neighbours distributed under

[7] http://espeak.sourceforge.net

[8] C stands for consonant, V stands for vowel and Y stands for semi-vowels [j], [ɥ] and [w].

the Lexique3 project, which includes 128,919 inflected forms. Based on findings in the cognitive psychology literature, we modelled this effect from different angles: the number of neighbours (11a), the cumulative frequency of all the neighbours (11b), and the number of more frequent neighbours (11c).

(13) *lexical frequency*: we used the lemma frequencies from Lexique3, which contains about 50,000 lemmas. Their frequencies were obtained from movie subtitles and were smoothed with the simple Good-Turing algorithm (Gale and Sampson, 1995) to assign a default frequency to out-of-vocabulary words. Preliminary experiments showed that it was better to use the logarithm of the frequencies, as commonly reported in the literature.

(15) *presence* in a list of simple words: a convenient proxy of the 'simplicity' of words. We then used a binary feature telling us whether this word is in the Gougenheim list (Gougenheim et al., 1964) or not. Since it was not obvious which size of list would be the best, we experimented with several sizes, ranging from 1,063 to 8,875 words.

We are currently testing the remaining variables:

(10) *phonological neighbourhood*: the number of words having a maximum number of phonemes in common (minimal series such as 'bain' [bɛ̃], 'main' [mɛ̃], 'pain' [pɛ̃], etc.). Our hypothesis is that the higher the number of neighbours, the easier the word.

(12) *abstract-concrete* and *imageability*: concrete words, as well as vocabulary from familiar contexts, would have a lower level of difficulty than abstract words.

(14) *size of the derivational family*: as shown by Schreuder and Baayen (1997) for visual word recognition, the bigger the family, the lower the difficulty a word would have as a result of proximity.

4.2.2 Analysis of the variable efficiency

In this section, we analyze how a simple lexicon (obtained from the Parkinson corpus) deviates, according to our variables, from general trends in the language, as represented by Lexique3.

For each variable listed in the previous section[9], we compared its distribution on both corpora using statistical tests. More precisely, a T-test (t) was applied to parametric interval variables, a Mann-Whitney test (U) to non-parametric interval variables, and a Chi-square test (X^2) to nominal variables (see Howell, 2008 for details). Figure 8 reports the means on both corpora (when meaningful) along with the p-values of the statistical tests.

[9] Presence in the Gougenheim list (15) was not considered for this step of the analysis, since this feature is not an intrinsic characteristic of words.

	Park.	**Lex3**	**p-value**[10]
1. # letters	6.3	8.6	< 0.001 (t)
2. # phonemes	4.7	6.8	< 0.001 (t)
3. # syllables	1.96	2.89	< 0.001 (t)
4. syll. struct.	/	/	0.6 (X^2)
5. sound-script	1.05	1.14	0.0004
6. # ortho.	0.75	0.96	0.007 (X^2)
11. #	3.88	1.31	< 0.001 (U)
13. frequencies	756.7	19.5	< 0.001 (t)

Figure 8: Variation in means from both corpora and significance of the difference between means.

The mean number of letters, phonemes and syllables is lower in our simple lexicon than in the language as represented by Lexique3. Words used by Parkinson speakers have, on average, 6.3 letters, 4.7 phonemes and 1.96 syllables; whereas words in Lexique3 have, on average, 8.6 letters, 6.8 phonemes, and 2.89 syllables. All three differences are significant, which is not surprising since these variables have been known for long in the readability literature as good proxies for the lexical complexity of a text.

Word frequency (13) is another feature that has proven useful for text readability measures. We also notice a significant difference ($p < 0.001$) between the frequencies of simple words, which are more frequent on average than the terms from Lexique3.

More innovative approaches of the lexicon difficulty include our variables based on the sound-script correspondences (5) and the difficulty of specific spelling patterns (6). Interestingly, both variables show significant differences between both lexicons. It appears that a simple lexicon contains significantly less complex correspondences between the sound and the written form. Also, simple words comprise fewer complex spelling patterns: 0.75 on average for simple words and 0.96 for the general lexicon.

Finally, simple words have significantly more orthographic neighbours (11) ($p < 0.001$). According to psycholinguistic literature (Andrews, 1997), this characteristic yields a facilitation effect in English, but not in French. Our result appears inconsistent with these experimental findings, but this is likely due to the fact that we did not control for the frequency of words. Since simpler words are also more frequent and shorter, they also tend to have more neighbours. It is worth noting that this type of inter-correlation between our variables is a well-known issue that must be taken care of when variables are combined within a statistical model, such as in Section 4.2.3.

[10] The threshold alpha used in this study is 0.05, which means that any lower p-value in this table represents a significant difference between the distributions in the Parkinson corpus and Lexique3.

To conclude this analysis, we have shown that all our variables, except the syllabic structure of words, have a different behavior on a simple lexicon and on the general vocabulary. This can be interpreted as a validation of their effectiveness in predicting the difficulty of terms. The next section further investigates these predictive abilities, using a lexicon of words annotated in terms of their complexity (i.e. Manulex).

4.2.3 The difficulty model

Having confirmed that most of our predictors can be used in order to discriminate between simple and complex words, we used Manulex as a gold standard to describe more precisely the relation existing between one of our variables and word difficulty. This relation, captured through a Spearman correlation[11], informs us how a given variable varies in relation to the three levels of difficulty in Manulex. This analysis precedes a more integrated approach, where all efficient variables are combined within a statistical model, which will also be used to assess the difficulty of words.

Name of the variables	Spearman corr.[12]
1. # of letters	0.27
2. # of phonemes	0.3
3. # of syllables	0.27
11a. # neighbours	-0.25
11b. cumulative freq. of neighbours	-0.25
13. word log-frequencies	**-0.51**
15. presence in the 5000 first words from the Gougenheim list	**-0.41**
6. complex spelling patterns (nasal)	0.08
6. complex spelling patterns (sum)	0.05

Figure 9: Spearman correlation for the most meaningful variables.

The total number of variables we tested amounts to 27 (including the variants described in Section 4.2.1). Correlations for the most efficient of them are reported in Figure 9. A positive correlation infers that the difficulty of words increases as the value of the variable increases (e.g. longer words tend to be more complex), whereas a negative correlation corresponds to the opposite relationship (e.g. complex words tend to be less frequent).

[11] Spearman correlation formula is described among others in Howell (2008). We did not use the Pearson correlation here, since some of our variables do not have a linear relationship with difficulty (e.g. those based on orthographic neighbours).

[12] Due to the large number of words in Manulex, all correlations reported in this table are significant at the level $p < 0.001$.

One should note that among the set of predictors which do not significantly correlate with word difficulty are our three classes of syllabic structures. This finding is consistent with our previous analysis on the Parkinson corpus. More surprisingly, spelling patterns and the difference between the oral and written forms do not account for much of the word difficulty. On the contrary, the two best predictors are the logarithm of word frequencies and the presence/absence from the 5,000 first words of the Gougenheim list.

As a result of this analysis, we selected a subset of nine predictors from our 27, which correspond to the best variables, as listed in Figure 9. These variables were combined using support vector machines (Boser et al., 1992) – generally abbreviated in SVM. It is a generalized linear classifier widely used in automatic classification[13].

We trained the final classifier on all Manulex words, but first estimated its performance on new words using a five-fold cross-validation approach. This consisted of splitting the data into five folds, training a model on four folds and testing it on the last fold. The accuracies thus obtained are averaged to yield an estimate of the mean accuracy of our model. It is also worth noting that SVMs require setting some parameters: the kernel used, the cost (C) and *gamma*. We opted for a radial basis function (RBF) kernel and explored by grid search a limited amount of combinations of values for C and *gamma*. The best model (with C = 1 and *gamma* = 0.5) attained a 62% classification accuracy.

Such a performance certainly leaves room for improvement, but should also be considered against the difficulty of the task. As we reported previously, Bormuth's (1966) study stressed the complexity of automatically predicting word difficulty. Moreover, our current model's accuracy is nearly twice as good as a random classification.

5. Results and discussion

Applying our lexicon difficulty model to JdM words absent from Manulex, we were finally able to produce a list of 17,870 graded words with graded synonyms, which stands as the first gold-standard list of French words to be used for language comprehension or production. The resource is available at:

http://cental.uclouvain.be/resyf

As the synonyms were extracted from a contributive lexical network, they correspond to the target word with a precision rate of 100%. However, some drawbacks can be identified for some lexical units, as a result of using word forms instead of senses.

[13] For an implementation of SVM available in Python, we relied on scikit-learn (Pedregosa et al., 2011).

5.1 Drawbacks requiring a more fine-grained study of the vocabulary

By and large, we have identified two kinds of issues:

a) Semantics

Polysemy and homonymy are not yet taken into account, neither is the difference between concrete and figurative senses. As a consequence, our resource assigns the same difficulty level to the various senses of a given word. For example, the word *renard* in French means 'fox' in a literal sense, but it also refers to an 'intelligent or smart attitude'. The list of synonyms for this word is the following one:

renard(1) *futé(1), malin(1) / goupil(2), canidé(1)*

(fox / smart / canid)

The two senses should be distinguished and should probably get a different difficulty score. The same applies to the word *hospitalier* ('related to hospitals' in a first sense, 'friendly and welcoming' in a second interpretation)[14].

Another problem with the synonyms obtained is the register or language level. Three levels could be defined: familiar or slang, current, formal. A tag indicating the appropriate language register should be added. To give an example, *policier* ('police officer') has two synonyms belonging to a familiar register (*flic* and *poulet*, corresponding to 'cop'). Whether the lexicon is used by someone affected by a language difficulty or by a machine for a lexical simplification task, such information on senses and register should be taken into account.

b) Compounds

In Manulex, compounds mostly belong to levels 2 or 3, for example:

papier-monnaie(3) *argent(1), billet(1)*

(paper money / money, bill)

homme-orchestre(2) *musicien(1)*

(band man / musician)

However, in some cases, the semantics of the target word can be obtained by the 'sum' of the senses of the word-forms integrating the compound word:

[14] Identifying the semantic structure of lexical units is a crucial issue in NLP. In future work we will follow existing proposals already defined in the literature, (Ploux & Victorri 1998) among others.

yéti(3) *abominable(2) homme(1) des neiges(1)*

(Yeti / abominable snowman)

These intuitive examples show the interest of investigating compounding and lexicalization mechanisms. In future work, we intend to evaluate how to automatically relate semantic compositionality or opacity (which are not trivial to measure) to word difficulty.

5.2 NLP for building specialized lexicons

As in many disciplines, the use of semi-automatic methods and specific software has become widespread over the last decades. Responsibility for key lexicographic tasks has been transferred from people to computers, especially for those tasks at which the computer excels, namely, counting, clustering, treating large amounts of data, extracting patterns, and identifying salient neighborhoods between words, etc.

Since the 1980s, lexicographers benefit from ever-growing volumes of data and either the collection or the analysis of such data has become largely streamlined (the 'drudgery' in the words of M. Rundell, 2009). Progress in computational linguistics has permitted a deeper investigation of the data, discriminating surface differences and highlighting more fine-grained representations at the morphological, syntactic or even semantic level (Grefenstette, 1998).

As mentioned in previous sections of this article, statistics computed by machines on large volumes of data have shown interesting results on determining how simple a word can be (frequency effect). However, we show in this paper that more sophisticated measures have to be considered and that NLP methods are useful for obtaining them. In a first step, basic linguistic treatments (tokenizing, lemmatizing and part of speech tagging) allow us to identify lexical units in corpora. Counting phonemes or letters, syllabification or on the consistency sound-script (difference between number of letters and phonemes) are simple tasks for a computer. More difficult tasks may imply the use of computational lexicons with structured information. To give an example, to obtain the number of morphemes, a list of affixes is required, as well as some linguistic knowledge on phonological alternations. Similarly, to identify senses on polysemic words, explicit linguistic knowledge has to be gathered on available resources and clustering heuristics have to be implemented to regroup senses. Lastly, as we have shown, the design of a language model is crucial to predict the level of difficulty of a word by combining and weighting the different predictors over large amounts of data.

Judging from these examples, computational linguistics enables the formalization of fine-grained linguistic phenomena which, in turn, provides a better comprehension of such phenomena. As a result, specialized lexicons with explicit information can be created, for human or automated usages in NLP tasks.

6. Conclusion

In this paper, we presented the first version of a French lexicon of synonyms graded with a tag indicating the level of difficulty (*ReSyf*). The data and the tags were obtained from existing resources and from a lexicon difficulty model based on a set of lexical measures. Such measures describe fine-grained intra-lexical features as well as some statistical or psycholinguistic properties of words.

Although we present preliminary work, our contribution demonstrates that natural language techniques can be used to create lexical resources with specific information (in this case, the difficulty levels) gathered and tested over different kinds of corpora.

Yet, there remain important aspects that have to be taken into consideration. We already mentioned that a more accurate sampling of the levels in Manulex is required to refine the gold-standard list. Ideally, a more precise training resource should be obtained through large scale subject testing. In addition, some variables that we introduced have yet to be implemented and integrated to our model. Finally, we also highlighted the importance of a more semantic-oriented approach to the lexicon complexity (as word forms are ambiguous).

To conclude, our future research will continue to focus on the identification of the features that make words easier for a given population class (in particular populations with language impairments) as well as on the automatic assessment word difficulty. We thus foresee a comparison of pedagogical data with pathological data to obtain deeper insights, while adapting the model to take the senses into account. Finally, we expect to use *ReSyf* in the context of automatic text simplification. The integration of a graded resource of synonyms indeed seems likely to impact the efficiency of such systems.

7. Acknowledgements

This project is partly financed by the Programme Hubert Curien (PHC) Tournesol 2013 (France-Fédération Wallonie-Bruxelles).

8. References

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval : Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, 4(4):439–461.

Biran O., Brody S. & Elhadad, N. (2011). Putting it simply: a context aware approach to lexical simplification. Proceedings of the 49th *Annual Meeting of the Association of Computational Linguistics (ACL 2011)*, pages 496-501. Portland, Oregon.

Boser, B. and Guyon, I. et Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Bormuth, J. (1966). Readability: A new approach. *Reading research quarterly*, 1(3):79–132.

Brysbaert, M., Lange, M. and Van Wijnendaele, I. (2000). The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition: Further evidence from the Dutch language. European Journal of Cognitive Psychology, 12(1):65–85.

Calzolari, N., Gurevych, I. and Kim, J. (2013). *The People's Web Meets NLP: Collaboratively Constructed Language Resources* annotated edition. Springer-Verlag Berlin and Heidelberg GmbH & Co. K.

Collins-Thomson, K. and Callan, J. (2004). A language modeling approach to predicting reading difficulty. Proceedings of *Human Language Technologies (HLT-NAACL)*, pages 193-200.

Dale, E. (1931). A comparison of two word lists. Educational Research Bulletin, 10(18): 484–489.

Dale, E. and Chall, J. (1948). A formula for predicting readability. *Educational research bulletin*, 27(1):11-28.

De Belder, J. and Deschacht, K. (2010). Lexical Simplification. Proceedings of the *1st International Conference on Interdisciplinary Research on Technology, Education and Communication* (ITEC 2010). Kortijk.

Ferrand, L. (2007). *Psychologie cognitive de la lecture*. De Boeck, Bruxelles. (ISBN-13 9782804159030).

François, T. (2012). Lexical and syntactic complexities: a difficulty model for automatic generation of language exercises in FFL. PhD thesis. Université Catholique de Louvain, Louvain-la-Neuve.

François, T. and Fairon, C. (2012). An "AI readability" formula for French as a foreign language. In Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP 2012), Jeju, 466-477.

Gala, N. and Lafourcade, M. (2011). NLP lexicons: innovative constructions and usages for machines and humans. Proceedings of Electronic Lexicography (E-Lex 2011). Bled (Eslovenia).

Gale, W. and Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237.

Gernsbacher, M. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. Journal of Experimental Psychology : General, 113(2):256–281.

Grefenstette, G. (1998). The Future of Linguistics and Lexicographers: Will there be Lexicographers in the Year 3000?, in Fontenelle et al. (Eds) Proceedings of the Eighth *EURALEX Congress*. Liege: University of Liege: 25-41. Reprinted in Fontenelle, T (Ed.*)Practical Lexicography: A Reader*. OUP 2008.

Gougenheim G. (1958). *Dictionnaire fondamental de la langue française,* Paris : Didier. (ISBN 2-208-00133-8).

Howes, D. and Solomon, R. (1951). Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, 41(40):1–4.

Lafourcade, M. (2007). Making people play for Lexical Acquisition. Proceedings of

the 7th *Symposium on Natural Language Processing (SNLP-2007*). Pattaya, Thailande, 8 pages.

Laufer, B. (1997). What's in a word that makes it hard or easy: Some intralexical factors that affect the learning of words. In S CHMITT, N. et M C C ARTHY, M., editors: *Vocabulary: Description, Acquisition and Pedagogy*, pages 140–155. Cambridge University Press, Cambridge.

Lété, B., Sprenger-Charolles, L. and Colé, P. (2004). Manulex: A grade-level lexical database from French elementary-school readers. Behavior Research Methods, Instruments & Computers, 36, 156-166.

Levelt, W. J. M. (1989). Speaking. Cambridge, MA: MIT.

Ogden, C. K. (1930). Basic English: A General Introduction with Rules and Grammar. Cambridge.

McNamara, P. (2010). Parkinson's Disease-Related Speech and Language Problems. Retrieved April 3, 2013, from http://parkinsons.about.com/od/signsandsymptomsofpd/a/speech_problems.htm

Morrisson, C. and Ellis, A. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 21(1):116–133.

New B., Pallier, C., Ferrand, L. and Matos R. (2005). Une base de données lexicales du français contemporain sur Internet: Lexique. *L'Année Pschologique*, 101, 447-462.

O'Regan, J. and Jacobs, A. (1992). Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1):185–197.

Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825-2830.

Pinto, S. and Ghio, A. and Teston, B. and Viallet, F. (2010) La dysarthrie au cours de la Maladie de Parkinson. Histoire naturelle de ses composantes: dysphonie, dysprosodie et dysarthrie. Revue Neurologique, vol. 166, no. 10. 2010, p. 800-810.

Ploux, S. and Victorri, B. (1998) Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. Traitement Automatique des Langues, vol. 39(1) :161-182.

Rundell, M. (2009). The road to automated lexicography: First banish the drudgery... then the drudges? In Proceedings of *eLexicography in the 21st Century Conference*, Louvain-la-Neuve, Université Catholique de Louvain.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings *of International Conference on New Methods in Language Processing*, volume 12. Manchester, UK.

Schreuder R. and Baayen R. H. (1997). How simplex complex words can be. *Journal of Memory and Language* 37, 118-139.

Taylor, W. (1953). Cloze procedure : A new tool for measuring readability. Journalism quarterly, 30(4):415 433.

University of Groningen (2011). Parkinson's disease undermines language processing. ScienceDaily. Retrieved April 3, 2013, from http://www.sciencedaily.com/releases/2011/02/11020262.htm