Towards French Text Simplification: where are we now?

Núria Gala
(Aix Marseille Université - LIF, France)

UFRGS, Porto Alegre, 7 Janeiro 2015

Who am I :

- **N. Gala**, Assistant Professor at Aix Marseille Université (since 2004), previously at Xerox Research Centre Europe (1999-2003)
- PhD in 2003, robust parsing, lexicalization and grammar development
- Computational linguist, interested in lexical studies and modelisation, lexical resources, e-lexicography, semantics, psycholinguistics
- Co-supervision of 3 PhD students in : parsing (A. Hamdi), lexicology and computational lexicography (J. Aznar), word sense disambiguation and simplification (M. Billami)

2012-2014, joint work with :

- **T. François**, L. Browers, C. Fairon (Université catholique de Louvain, Belgium)
- D. Bernhard, A. Todirascu (Université de Strasbourg, France)
- M. Billami (Aix Marseille Université - LIF, France)

Readability and text simplification

- Readability : The sum total (including the interactions) of all those elements within a given piece of printed material that affect the success of a group of readers have with it. The success is the extent to which they understand it, read it at a optimal speed, and find it interesting. [Dale and Chall, 1948a]

- Simplification : The process of transforming a text into an equivalent which is more understandable by a target audience. [Saggion, 2013]

Readability

- First readability formulae based on statistics : linear regression with two variables (a lexical and a syntactic one), eg. [Flesch, 1948], [Dale and Chall, 1948b]
- Coginitive approach : providing evidence of cognitive (inference) and text structure factors (coherence, cohesion)
- Computational approach : integrates both paradigms = based on automatic extraction of a set of different variables, use of statistical algorithms (SVM) to classify texts

Text Simplification

- Increasing interest in NLP but also in applied fields
- Explicit needs for particular users facing difficulties in learning vocabulary, reading and comprehending texts
  - Learning languages (L1 or L2) [Brouwers et al., 2014], deaf writing
  - People with written or spoken language impairments (dyslexia [Rello et al., 2013], aphasia [Watanabe et al., 2009])
  - Illiterates (low level of instruction) [Carroll et al., 1998], [Inui et al., 2003]
- Real scientific challenge of tremendous societal relevance
- Aims : to propose simplified texts that convey the same meaning to help to read better and to read more

Outline

First experiments on administrative texts

T. François

- Administrative issuances : difficult to be understood by a significant part of the population
- Aims : to propose a readability formula to categorize administrative text from 1 (very easy) to 5 (very difficult).
- Main problem : obtaining a training corpus
  - Works on computational readability = mainly learning material, already classified by the creators of that material
  - Nothing exist for administrative texts

How to anotate ?

Reading speed and expert judgement

- 115 real texts (Belgium) digitalized (XML) and splitted into 220 samples.
- Evaluation of the difficulty with a formula from [Kandel, 1958] (to ensure representative data)
- 10 texts with different levels of difficulty tested with AMesure-Testing $\rightarrow$ annotation guide
- Manual annotation by experts from belgian administration (α de Kripendorf = 0.37).

At the end, 115 annotated samples, 5 levels of difficulty

Reading speed average per text

Document	KM score	ms. /word	Level
La santé de votre enfant	71.3	292.8	1
Du couple à la famille	86.5	304.9	1
Des chaussures... Quand les mettre aux pieds ?	81.1	315	2
A l'école d'une alimentation saine	75.8	324.4	2
L'enseignement spécialisé	46.2	339.7	3
Lettre pour la semaine européenne de la vaccination	40.6	340.5	3
Cumuls de pensions	57.5	372.3	4
Liquidation des subventions ordinaires 2004	15	376.6	4
Déclaration de succession	57	379	5
Tax shelter	36.5	390	5

Formula (AMesure)

- Variables identified from **T. François** PhD work [François, 2011], among the most efficient :
  - average of words per sentence ($r > 0,64$)
  - ratio of pronouns and conjonctions, proportion of words > 8 letters, cumulated frequency of orthographical neighbours, personnalisation rate of texts, proportion of past participles, coherence inter-sentence, etc.
- Model : Support Vector Machine (SVM), selection of variables (predictors) based on a correlation analysis.

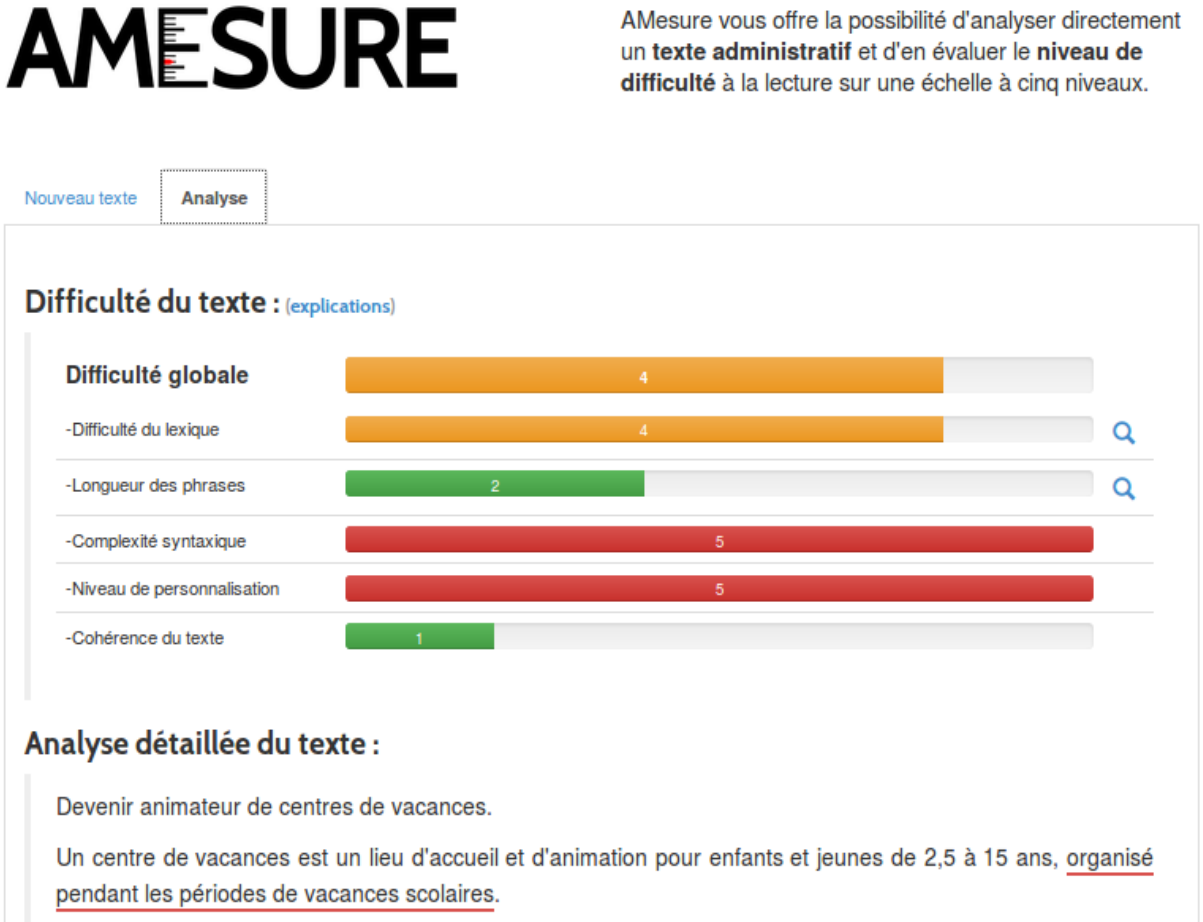
  $\rightarrow$ Model based on 11 variables : $acc = 50\%$ and $adj - acc = 86\%$ for 5 levels.

On line Plateform

Formula integrated in AMesure :

- Lexical complexity :
  - based on frequencies from Lexique 3 (a standard French lexical database) [New, 2006]
- Syntactic complexity :
  - Based on preliminar work on syntactic typology by **L. Brouwers** [Brouwers et al., 2014]
  - Manual analysis from parallel corpus (original and simplified, Wikipedia and Wikimini)
  - 19 rules for identifying passive sentences, subordinate clauses and parenthesis.

Demo AMesure



http ://cental.uclouvain.be/amesure/

Graded Lexicons : overview

Lexical repositories where words have a level of difficulty associated to them, calculated upon different variables (statistic measures but also linguistic features).

- *Teachers' Book of Words* [Thorndike, 1921]
- *Français fondamental* [Gougenheim, 1958]
- ...
- *Manulex* [Lété et al., 2004] > French (L1)
- *ReSyF* [Gala et al., 2013] > French (L1)
- *FLELex* [François et al., 2014] > French (L2)
  - based on CEFR (*Common European Framework of Reference for Languages*)
  - 6 levels : A1/A2 (beginners), B1/B2 (intermediate), C1/C2 (advanced)

Fundamental Lexicons : forerunner idea

List of minimal vocabulary of a language (lists of 'easy' words).

- New approaches for teaching languages (begin XXth century)
- Controversial idea > simplify the vocabulary = reductionist
- Applied to readability formulae
- Rationnal approaches (humans), eg. *Basic English* [Ogden, 1930]
- Statistical approaches (corpus), eg. *Teacher's Book of Words* [Thorndike, 1921]
- Hybrid approaches, frequencies obtained for familiar words, eg. *Français Fondamental* [Gougenheim, 1958], *Les listes orthographiques de base du français* [Catach, 1984]

Drawbacks of firsts lists with simplified words

- Based on a single criterion : word frequencies
- Limited coverage (a few of hundreds easy words : 850 Ogden, 1,500 Gougenheim, 30,000 Thorndike)
- Early computational approaches, words in contexts [Bormuth, 1966] :
  - multiple correlation coefficient (R) with four variables (number of syllables, number of letters, frequency index, word depth)

Predicting difficulty of words is surprisingly harder (R = 0.505) than predicting text difficulty (R = 0.934)

Manulex

[Lété et al., 2004]

- School graded list of words, 19 037 base-forms (lemmas)
- Three levels (1st year, 2nd year, three following years of primary school)
- Several frequency measures (raw frq, dispersion index, etc.)
- Transforming such measures into a class : attested presence in a school level

Word	POS	Level 1	Level 2	Level 3
pomme *(apple)*	N	724	306	224
vieillard *(old man)*	N	-	13	68
patriarche *(patriarch)*	N	-	-	1
cambrioleur *(burglar)*	N	2	-	33
Total in Manulex		31%	21%	48%

- Building our first gold-standard list of graded words

ReSyf

[Gala et al., 2013]

- Resource with graded synonyms
- Automatically built from different existing resources :
  - French Lexicons : *Lexique 3* [New et al., 2001], *Manulex* [Lété et al., 2004], *JeuxDeMots* [Lafourcade, 2007]
  - Corpus : sample of a corpus from patients with Parkinson (lexical analysis of 1 106 lemmas)
- Initial list : 19 037 lemmas from Manulex (Adjs, Advs, Nouns, Verbs), transformation in 3 classes depending on their presence in a level in Manulex
- Final list (version 2013) : 12 687 lemmas lexicaux de Manulex with synonyms in the lexical network JeuxDeMots

Data from ReSyf

Graded synonyms FR L1

renard(n1) : malin(n1) futé(n1) goupil(n2) canidé(n3) roublard(n3)
pourtant(n1) : cependant(n1) néanmoins(n2) seulement(n1) toutefois(n2)
armure(n1) : cuirasse(n2) tissage(n3) harnais(n3) protection(n1)
piétiner(n2) : fouler(n3) piaffer(n3) trépigner(n1) marcher(n1)
glacial(n2) : sec(n1) froid(n1) insensible(n3) glacé(n1) polaire(n2) impassible (n3) imperturbable(n3) rigoureux(n2) inhospitalier(n3)
patriarche(n3) : chef(n1) vieillard(n2) père(n1)
joncher(n3) : couvrir(n1) parsemer(n2) tapisser(n1) disséminer(n3) recouvrir(n1)
policier(n1) : poulet(n1), flic(n2), commissaire(n3)
extravagance(n3) : absurdité(n3) folie(n1) bizarrerie (n3) frasque(n2) caprice(n1) excentricité(n3) originalité(n3) démence(n3) fantaisie(n2)

Work in progress : new version of ReSyf with higher coverage and word sense disambiguation (M. Billami PhD)

FLELex

[François et al., 2014]

- List of words created from CEFR corpora (777 835 words, FR L2)
- Extraction of 16 833 lemmas lexical words (1 038 grammatical words)
- Tokenization and tagging with NLP tools (Treetagger and CRF tagger)
- 31% of words with frequency > 10 (6% freq > 100) and 69% < 10 (with 20% hapax)
- Comparison to standard French in Lexique 3 (47 342 lemmas) : 622 words from FLELex are absents (3,5%)

Corpus

28 FLE books and 29 simplified books (777 835 words) :

Genre	A1	A2	B1	B2	C1	C2	Total
Dialogue	153 (23,276)	72 (17,990)	39 (11,140)	5 (1,698)	/	/	269 (54,104)
E-mail, mail	41 (4,547)	24 (2,868)	44 (11,193)	18 (4,193)	8 (2,144)	1 (398)	136 (25,343)
Sentences	56 (7,072)	21 (4,130)	12 (1,913)	5 (928)	/	/	94 (14,043)
Varias	31 (3,990)	36 (4,439)	23 (5,124)	14 (1,868)	1 (272)	/	105 (15,693)
Text	171 (23,707)	325 (65,690)	563 (147,603)	156 (63,014)	175 (89,911)	48 (34,084)	1,438 (424,009)
Readers	8 (41,018)	9 (71,563)	7 (73,011)	5 (59,051)	/	/	29 (244,643)
Total	460 (103,610)	487 (166,680)	688 (249,984)	203 (130,752)	184 (92,327)	49 (34,482)	2,071 (777,835)

Dispersion index

- The corpus has been tagged with two taggers : TreeTagger and CRF tagger (MWEs detection).
- A dispersion index for each word has been calculated :

$$D_{w,K} = \log(\sum p_i) - \frac{\sum p_i \log(p_i)}{\sum p_i}] / \log(I) \quad (1)$$

With K = level ; I = number of books for the level K ; p_i = the probability of appearence of a word in the book i.

- Frequencies have been modified by D and they have been normalized :

$$U = (\frac{1\ 000\ 000}{N_k})[RFL * D + (1 - D) * f_{min}] \quad (2)$$

With N_k = number of tokens in the level k and $f_{min} = \frac{1}{N} \sum f_i s_i$ with f_i = the frequency of a word in the book i and s_i = number of words in the book i

FLELex Data

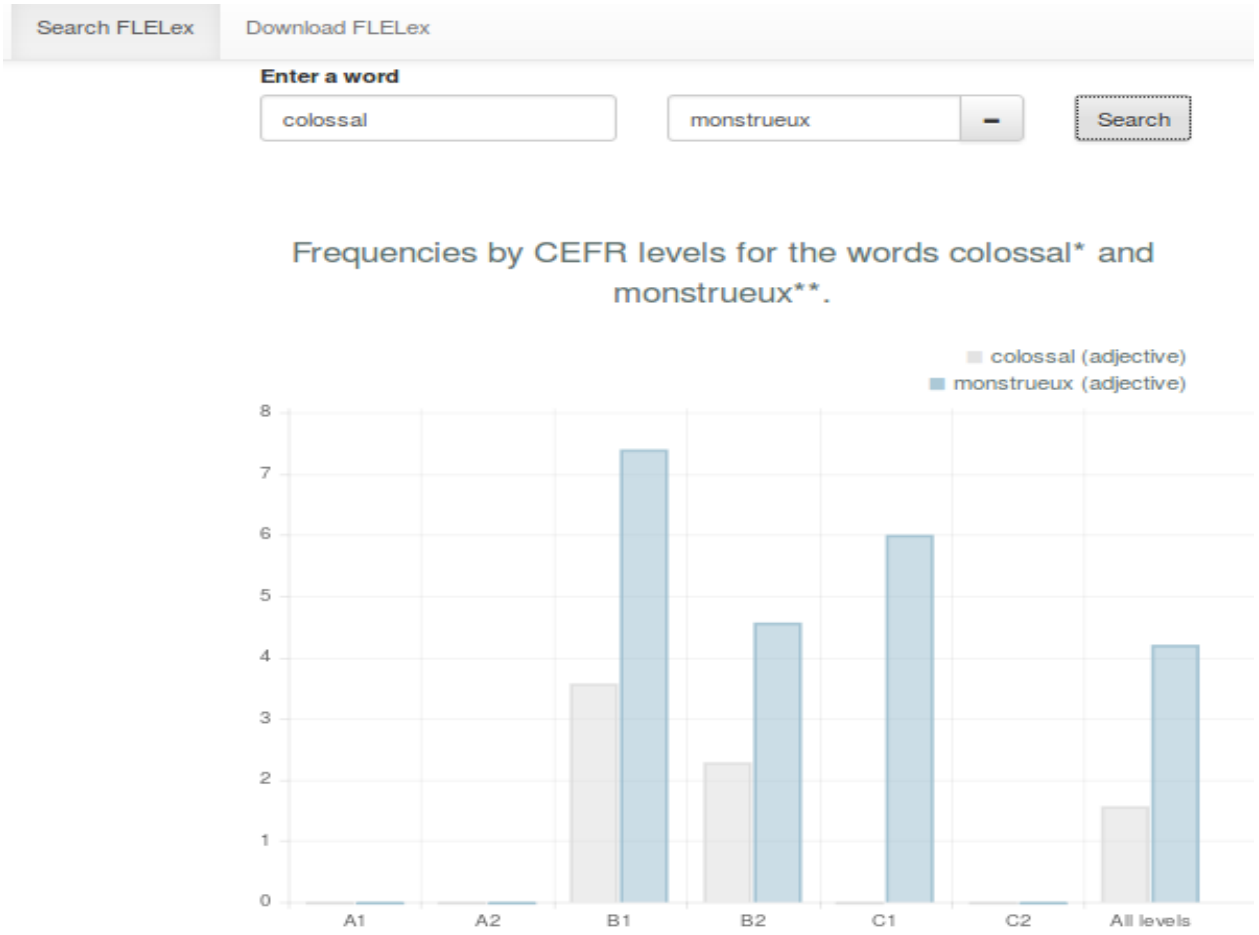
- List of 17 871 lemmas.

lemme	A1	A2	B1	B2	C1	C2
voiture	633.3	598.5	482.7	202.7	271.9	25.9
abandonner	35.5	62.3	104.8	79.8	73.6	28.5
justice	3.9	17.3	79.1	13.2	106.3	72.9
kilo	40.3	29.9	10.2	0.0	1.6	0.0
piétiner	0.0	0.39	0.0	0.53	15.7	0.0
logique	0.0	0.0	6.8	18.6	36.3	9.6
absurdité	0.0	0.0	0.34	4.55	3.29	67.36
en bas	34.9	28.5	13	32.8	1.6	0.0
en clair	0.0	0.0	0.0	0.0	8.2	19.5
de surcroît	0.0	0.0	0.0	0.0	15.67	0.0
donner rendez-vous	0.53	0.69	1.89	0.0	0.0	0.0
donner naissance	0.0	0.25	0.0	0.0	0.0	4.12

Initial levels : more familiar and concrete words

Higher levels : more literary and abstract words

Demo FLELex



http ://cental.uclouvain.be/flelex/

Corpora used in our experiments

1. PK_corpus : a small corpus of parkinsonian patients
2. LL_corpus : language learning at school (French tales and fables, short stories to learn to read at school)

(1) Language impairments : Parkinsonian corpus

- Parkinson disease : motor symptoms but also language and speech impairments (hypophonia, monotone speech, difficulties in articulation) [Pinto et al., 2010]
- 20 recordings of patients in 'off state', 2 271 tokens (occurrences), 1 106 base-forms (lemmas Adjs, Advs, Nouns, Verbs)
- Guided task : to describe a picture
- Average lenghts of words :
  - Corpus Pk (Parkinson) : 6,3 letters, 4,7 phonemes, 1,96 syllabes
  - Lexique 3 (standard French) : 8,6, letters, 6,8 phonemes, 2,9 syllabes
- Levels according to Manulex :

	Level 1	**Level 2**	**Level 3**
Total corpus Pk	94,3%	1,45%	1,63%

Parkinsonian corpus from patients describing a picture

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE Trans SYSTEM "trans-14.dtd">
<Trans scribe="Nuria" audio_filename="A02403_off" version="2" version_date="130315">
<Episode>
<Section type="report" startTime="0" endTime="48.906">
<Turn startTime="0" endTime="48.906">
<Sync time="0"/>
ah je vois un petit qui est en train de se casser la gueule il est en train de bouffer des
gâteaux
<Sync time="6.875"/>
y a sa y a sa soeur qui en voulait un peu mais enfin
<Sync time="11.115"/>
et là il y a la la femme qui est en train de laver sa sa vaisselle qui vient
<Sync time="19.431"/>
(...) oh il note
<Sync time="23.652"/>
oh le garçon il a un très court un pantalon
<Sync time="30.734"/>
la fille a une robe des chaussures
<Sync time="33.757"/>
la femme elle a une serviette à la main (...) blanche
<Sync time="41.108"/>
y a deux tasses parce qu'elle les a lavées
</Turn>
</Section>
</Episode>
</Trans>
```

(2) Language learning : corpus used at school and reading tests

- French L1, five years of primary school : children from 6 to 11 years old
- Manual simplifications :

- vocabulary : lexical removing or replacements
- morphology : morphological replacements (tense, words with irregular inflection)
- syntax : sentence structure (subordinates, passive, negation, reformulations)
- discourse : pronoun specification (anaphora removing)

- Right now 18 corpus 2 700 tokens (average 150 tokens original, 144 tokens simplified version)
- Annotation (work in progress) up to 76 corpus (11 400 tokens)

Corpus Sample

L'hiver, les animaux et les plantes vivant dans les régions froides doivent trouver des moyens pour subsister.

Beaucoup de plantes hibernent sous forme de graines qui germent au printemps pour se transformer ensuite en de nouvelles plantes.

Il y en a d'autres dont la partie apparente meurt et, quand la température se réchauffe, elles créent de nouvelles pousses.

En automne, beaucoup d'arbres perdent leurs feuilles et durant l'hiver, ils observent une période de repos.

Quant aux animaux, ils consomment beaucoup plus d'énergie que les plantes car ils bougent.

La plupart des animaux survivent l'hiver sans changer leur façon de vivre habituelle.

Mais il existe une catégorie d'animaux qui doit prendre des dispositions particulières pour ne pas mourir de froid

L'hiver, les animaux et les plantes vivent dans des régions froides et doivent trouver des moyens pour survivre.

Beaucoup de plantes restent en vie l'hiver sous forme de graines qui grandissent au printemps pour se transformer ensuite en de nouvelles plantes.

Il y d'autres plantes qui perdent leur partie visible et, quand la température se réchauffe, elles fabriquent de nouvelles branches.

En automne, beaucoup d'arbres perdent leurs feuilles et pendant l'hiver, ils ne font rien.

Les animaux utilisent beaucoup plus d'énergie que les plantes parce qu'ils bougent.

Beaucoup d'animaux passent l'hiver sans changer leur façon de vivre.

Mais il existe un type d'animal qui doit changer ses habitudes pour ne pas mourir de

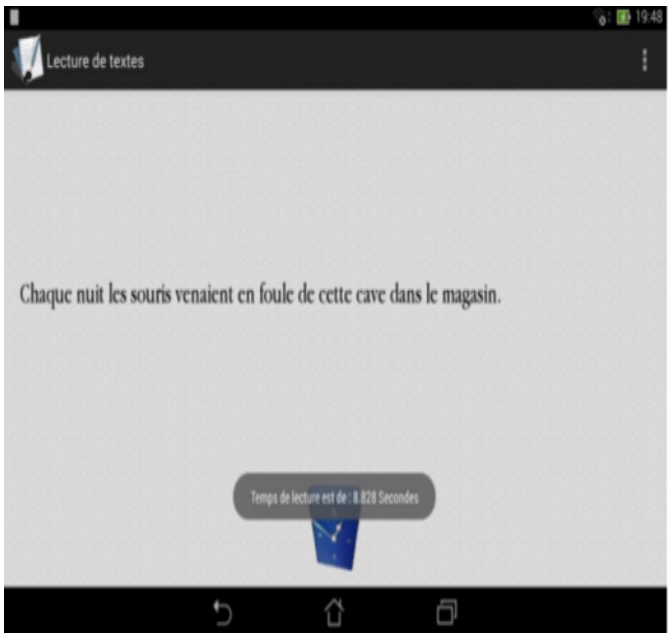
Using the lexicons and the corpora for experiments

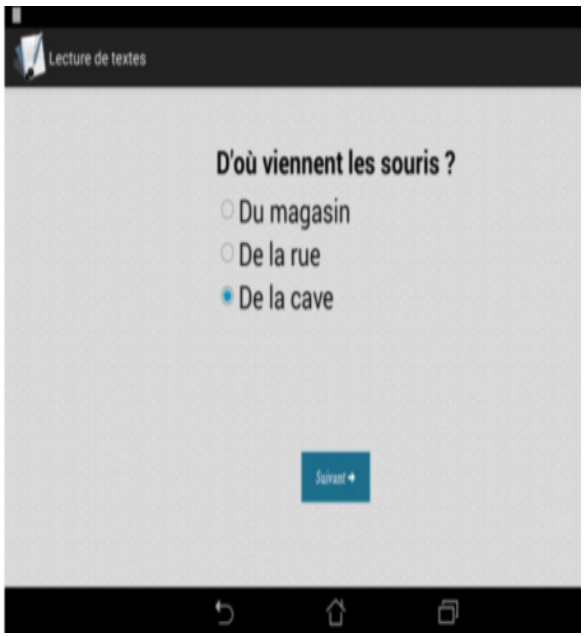
- Lexicons : variables for the model on lexical complexity
- Parkinson corpora : variables for the model on lexical complexity

- Language learning corpora, school materials (work in progress) :
  - testing with children with dyslexia and poor readers : reading speed and comprehension questions
  - studying the results obtained to better model a text simplification tool for dyslexic readers and poor readers

Reading Test

M. Billami





Lexical complexity : how complex a word is ? for whom ?

- Complexity :
  - Subjective notion [Blache, 2011],
  - Difficulty in reading and/or comprehending given a target audience
- Key idea :
  - Strong correlation between frequency and difficulty [Brysbaert et al., 2000]
  - Statistical factors (word-form frequencies)
- But other psycholinguistic factors (related to the 'perception' a user has) :
  - familiarity,
  - age of acquisition of vocabulary,
  - orthographical neighbours,
  - etc.

Lexical complexity : how complex a word is ? for whom ?

- What about the impact of linguistic factors ?
- Intralexical features inherent to each language :
  - syllable structure,
  - grapheme-phoneme consistence,
  - flexional and derivational regularities,
  - number of morphemes in a word [Schreuder and Baayen, 1997],
  - polysemy [Laufer, 1997]),
  - ...

Identifying lexical complexity (1/4)

[Gala et al., 2014]

- Identify the predictors of lexical complexity
- Use them to automatically predict the level of difficulty of a word
- Adapt them to the needs of a target audience (right now, L1 and L2 learners)
- Statistical and intralexical variables extracted from different resources
  - Graded lexicons : Manulex, FLELex
  - General lexicons : Lexique3, Morphalou, Polymots
  - Semantic networks : JeuxDeMots, BabelNet

Identifying lexical complexity (2/4)

Resources : Lexique 3, Manulex

- Number of letters, phonemes, syllabes
- Syllabic structure (more frequent structures in French V, CVC, CV)
- Grapheme-phoneme consistence :
  - 0 = transparence : 'abruti' [abryti]
  - $<$ 2 characters : 'abrite*r*' [abrite]
  - $>$ 2 characters : 'l*ent*em*ent*' [lãtmã]
- Orthographical patterns :
  - double vowels (eg. ée [e]),
  - double consonnants (eg. pp [p]),
  - digraphs (eg. ch [ʃ]),
  - nasal vowels (eg. -am/-an/-em/-ent [ã], -in/-ain [ɛ̃], -on [õ], -um/-un [œ̃])

Identifying lexical complexity (3/4)

Resources : Morphalou, Manulex, Polymots, FLELex

- Orthographic neighbours (the higher the number of similar words = the easier the word)
- Morphemes :
  - non supervised morphological analysis, split the words into tagged morphemic segments (root, prefix, suffix, linking element) and identify morphological families [Bernhard, 2010]
  - number of morphemes, prefixation (yes/no), suffixation (yes/no), compound (yes/no), minimal frequency pref/suf, average frequency pref/suf, morphological family size

rouille – antirouille ; rouilleux
dérouiller – dérouillage ; dérouillement ;
débrouille – brouilleur ; brouilleuse ; débrouilleur ; débrouilleuse
brouille – brouillerie ; brouilleux

Identifying lexical complexity (4/4)

Resources : JeuxDeMots, BabelNet

- Polysemy :
  - Two lexico-semantic networks available for French
  - *JeuxDeMots* (`http://www.jeuxdemots.org`) [Lafourcade, 2007] > has more than one sense (yes/no)
  - *BabelNet* (`http://babelnet.org/`) [Navigli and Ponzetto, 2010] > number of synsets (synonym groups)

rouille(r_infopot#36 :25–> _INFO-POLYSEMIC) [’altération’, ’rubigineux’, ’sauce’, ’érosion’]
rouille(3)
///bn :00068634n|noun|rouille///bn :00068636n|noun|rouille///bn :00068637n|noun|champignon

Results (1/3)

T. François and **D. Bernhard**

- Effectiveness of each variable evaluated alone with a Spearman correlation, using two different lexicons : Manulex (L1) and FleLex (L2)
- 49 variables identified :
  - Presence/absence in Gougenheim lists
  - Number of letters, phonemes, syllabes, phoneme-grapheme consistence, syllabic structure
  - Orthographic neighbourhood, orthographical patterns (double vowels, doble consonnants, nasal vowels, digraphs etc.)
  - Number of morphemes, prefixation, suffixation, frequence of affixes, compounds, morphological family size
  - Polysemy (from JeuxDeMots and BabelNet)

Results (2/3)

Variables	Manulex (ρ)	FLELex (ρ)
17) Fréquences dans Lexique3	-0,51	-0.53
18) Percentage absents Goug. list (5000)	-0,41	-0.46
18) Percentage absents Goug. list (4000)	-0,41	-0.47
02) Number phonemes	0,30	0,27
15) Polysemy JeuxDeMots	-0,29	-0.38
01) Number letters	0,27	0,25
03) Number syllables	0,27	0,26
4a) Number orth. neighbours	-0,25	-0,23
4b) Neighbours (cumulated frequency)	-0,25	-0,23
16) Synset BabelNet	-0,20	-0,19
6b) Nasal Vowel	0,08	0,07
14) Morphological Family size (morphoclust_10)	-0,08	-0,05
11) Prefixation (seg_10)	0.07	0,06
08) Number de morphemes (seg_10)	0,06	0,08
06) Orthographical patterns (a-d)	0,05	0,06
10) Suffix average (freq_seg_10)	-0,05	0,02

Results (3/3)

- Training a machine learning algorith SVM using the best edictors
- Evaluation of the model usinf Manulex (26 variables) and FLELex (24 variables)
- Baseline 1 : predicting the majority class
- Baseline 2 : model based only on frequencies of words

Liste	Modèle	Cost	Accuracy	Standard deviation
Manulex	Majority class	/	48%	/
	Word frequencies	0,1	61%	0,4%
	Our model	0,5	63%	0,7%
FLELex	Majority class	/	28,8%	/
	Word frequencies	0,5	39%	0,8%
	Our model	0,001	43%	0,5%

TABLE : Performances of the models on Manulex and FLELex.

Discussion (1/2)

- Hypothesis : imbalance of classes entails a 'ceiling effect' in the performances
  $\longrightarrow$ new experiences on a more balanced corpus (classes)

	All data (19 037)	**Sample (11 993)**	
		Same distrib.	**Uniform distrib.**
frequency	61,2%	61,6%	51,2%
selected variables	62,7%	63,0%	53,7%
all variables	62,9%	63,0%	53,4%

- No impact on the size but lower performances on balanced classes.

Discussion (2/2)

The model better classifies the words mainly on the extremes.

	All data (14 364 instances)	Sample (2 872 instances)	
		Same distrib.	Uniform distribution
A1 (4 142)	0,628	0,631	0,572
A2 (2 735)	0,029	0,041	0,326
B1 (4 002)	0,492	0,493	0,123
B2 (1 312)	0,000	0,023	0,251
C1 (1 672)	0,089	0,098	0,298
C2 (501)	0,000	0,000	0,352

TABLE : F-Mesure / level on FLELex.

Conclusions about our model

- A model for predicting lexical complexity (supervised learning)
- Significant number of statistical and intralexical variables
- As reported in the litterature, statistical variables are the best predictors
- Intralexical predictors bring a slight improvement
- Hypothesis :
  - Need to better balance de training data (number of words/level)
  - Extremes are better classified
  - The choice of levels in both Manulex (L1) and FLElex (L2) is done on a subjective and educational criteria
  - Need of new tests with population (eg. dyslexic children) to obtain psycholinguistic predictors

Conclusions and future work (1/2)

- Studies on readability and lexical complexity for French
- Generation of language resources :
  - Graded lexicons > FLELex (2014) and ReSyF 2.0 (2015)
  - Simplified corpora > parallel corpora, children litterature (2015)
- Tests on children with dyslexia and poor readers, psycholinguistic analysis and typology of phenomena (2015) > in collaboration with students (speech therapists) and prof. J. Ziegler, Laboratoire de Psychologie Cognitive (LPC)
- Annotation guide, linguistic and statistical analysis, comparisons O/S (2015-2016)

Conclusions and future work (2/2)

- Towards a readibility model for different target audiences and text types
- Use the results of linguistic analysis and experiments with population to develop a tool for text simplification in French
- Integrate cognitive and discursive measures [Todirascu et al., 2013], build an annotated corpus with anaphoric links
- Target other audiences (deaf people, other specific language impairments)

Collaborations with UFRGS and PUC

Some first ideas...

- Collect and analyze texts for illiterates in FR, multilingual comparisons (PT-FR) > in collaboration with Bianca Pasqualini
- Improve lexical resources with better visualization tools > in collaboration with Lucelene Lopes

- Study lexical complexity from a multilingual point of view, integrate findings from the different teams
- Develop a multilingual simplification tool (EN, PT, FR), focused on lexical simplification, integrate multi-word expressions (MWEs) and word sense disambiguation (WSD)

Muito Obrigada !

Núria Gala
Aix Marseille Université - LIF (France)

nuria.gala@lif.univ-mrs.fr


Bernhard, D. (2010).
Apprentissage non supervisé de familles morphologiques : comparaison de méthodes et aspects multilingues.
Traitement Automatique des Langues, 2(51) :pp. 11–39.


Blache, P. (2011).
A computational model for language complexity.
In *1st Conference on Linguistics, Biology and Computational Science*, Tarragona, Spain.


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


Brouwers, L., Bernhard, D., Ligozat, A.-L., and François, T. (2014).
Syntactic sentence simplification for french.
In *Proceedings of the 3rd International Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014)*.


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.


Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998).
Practical simplification of English newspaper text to assist aphasic readers.
In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*.


Catach, N. (1984).
Les listes orthographiques de base du français.
Nathan, Paris.


Dale, E. and Chall, J. (1948a).
The concept of readability.
Elementary English, 26(1) :19–26.


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


Flesch, R. (1948).
A new readability yardstick.
Journal of Applied Psychology, 32(3) :221–233.


François, T. (2011).
Les apports du traitement automatique du langage à la lisibilité du français langue étrangère.
PhD thesis, Université Catholique de Louvain.


François, T., Gala, N., Watrin, P., and Fairon, C. (2014).
FLELex : a graded lexical resource for French foreign learners.
In *Proceedings of International conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.


Gala, N., François, T., and Fairon, C. (2013).
Towards a French lexicon with difficulty measures : NLP helping to bridge the gap between traditional dictionaries and specialized lexicons.
In *E-lexicography in the 21st century : thinking outside the paper.*, Tallin, Estonia.


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

Un modèle pour prédire la complexité lexicale et graduer les mots.
In *Actes de TALN 2014*, Marseille.


Gougenheim, G. (1958).
Dictionnaire fondamental de la langue française.
Didier, Paris.


Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003).
Text simplification for reading assistance : a project note.
In *Proceedings of the second international workshop on Paraphrasing*, pages 9–16.


Kandel, L. et Moles, A. (1958).
Application de l'indice de flech à la langue française.
Cahiers Etudes de Radio-Télévision, 19 :253–274.


Lafourcade, M. (2007).
Making people play for Lexical Acquisition.
In *Proc. SNLP 2007, 7th Symposium on Natural Language Processing.*, Pattaya, Thailande.


Laufer, B. (1997).

What's in a word that makes it hard or easy : Some intralexical factors that affect the learning of words.
Cambridge University Press.


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


Navigli, R. and Ponzetto, S. P. (2010).
BabelNet : building a very large multilingual semantic network.
In *48th annual meeting of the Association for Computational Llinguistics.*, pages 216–225, Uppsala, Suède.


New, B. (2006).
Lexique3 : une nouvelle base de données lexicales.
In *Actes de TALN 2006*, Louvain la Neuve, Belgium.


New, G. A., Pallier, C., Ferrand, L., and Matos, R. (2001).
Une base de données lexicales du français contemporain sur Internet : Lexique 3.

L'année psychologique, 101 :447–462.

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

Pinto, S., Ghio, A., 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, 166(10) :800–810.

Rello, L., Baeza-Yates, R., and Saggion, H. (2013).
The impact of lexical simplification by verbal paraphrases for people with and without Dyslexia.
Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science, 7817 :501–512.

Saggion, H. (2013).
Automatic text simplification : What for ? invited speaker at recent advances in natural language processing (ranlp).

Schreuder, R. and Baayen, H. (1997).

How complex simplex words can be.
Journal of Memory and Language, pages 118–139.


Thorndike, E. (1921).
The Teacher's Word Book.
Teachers College, New York.


Todirascu, A., François, T., Gala, N., Ligozat, A.-L., and Bernhard, D. (2013).
Coherence and Cohesion for the Assessment of Text Readability.
In *Proceedings of the 10th International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2013)*, Marseille.


Watanabe, W. M., Junior, A. C., Uzêda, V. R., Fortes, R. P. d. M., Pardo, T. A. S., and Alusio, S. M. (2009).
Facilita : reading assistance for low-literacy readers.
In *SIGDOC '09 : Proceedings of the 27th ACM international conference on Design of communication*, pages 29–36, New York, NY, US.