Assessing the lexical complexity in French: FLELex and ReSyf

Thomas François

Seminar in KULeuven

december 17, 2014

Plan

Plan

The challenge of reading

Reading remains a challenge for a significant part of the population, even in our highly educated societies :

- UE recent report (2009) : 19, 6% of 15-year teenagers are “low achievers” [De Coster et al., 2011, 22]
- [Richard et al., 1993] : On 92 unemployement benefit form filled by people with a low education level, half of the required information was missing.
- [Patel et al., 2002] : Their subjects faced significant problems in understanding the different steps for the proper administration of drugs.
- Besides, reading is also an issue for the large amount of L2 learners faced with written texts (in lecture, administration, web, etc.)

Reading and NLP

Natural language processing can help low readers in various ways :

- Automatic selection of reading materials at their level (readability) ;
- Automatic generation of reading or language exercices ;
- Integration within iCALL software for intelligent feedback, better adaptability or incremental content collection ;
- Automatic text simplification (ATS) to improve access to of authentic texts ;
- Difficulty diagnosis of texts for writers.

An example : AMesure

AMesure is a free web platform that assess the difficulty of administrative texts :

- Includes a readability formula that classifies texts on a 1-to-5 scale ;
- Trained on a small corpus of 115 texts (annotated by FWB experts) ;
- Selection of 11 variables among 344 : model reaches $acc = 50\%$ and $adj - acc = 86\%$;
- Besides the formula, lexical and syntactic diagnosis is provided.

The issue of vocabulary

Vocabulary and L2 learning

- Vocabulary knowledge is crucial for L2 learning and a reader must know between 95% to 98% of the words in a text to adequately understand it [Hu and Nation, 2000]
- In readability formulas, the lexical features have been shown to account the most for text difficulty [Chall and Dale, 1995]

- Control the level of vocabulary in a text is therefore valuable for learning...
- It can also be useful for other tasks, such as text simplification.

In this talk, we aim at assessing the difficulty of the lexicon

Assessing lexicon difficulty

Psycholinguistic investigates the complexity of words through various dimensions :

- Word frequency effect : correlation between frequency of words and difficulty [Brysbaert et al., 2000]
- The age-of-acquisition seems to play a role in decoding, independently of the word frequency [Gerhand and Barry, 1999]
- The number of orthographic neighbours [Andrews, 1997]
- Concretedness and imageability of words [Schwanenflugel et al., 1988]
- The familiarity of readers with words (and morphemes) also helps recognition [Gernsbacher, 1984]
- The number of (known) senses [Millis and Button, 1989]

Approaches in L2 learning and teaching

- There is also a bunch of studies in vocabulary learning that correlates words characteristics with ease of learning.
- [Laufer, 1997] focused on factors such as familiarity of phonemes, regularity in pronunciation, fixed stress, consistency of the sound-script relationship, derivational regularity, morphological transparency, number of meanings, etc.

- Another approach is to defined gradued lexicon lists on which the learning process and materials selection can be based.
  $\rightarrow$ Question : how are these lists obtained ?

Frequency lists

- One of the first lists was collected by [Thorndike, 1921] : list of 10,000 words with frequencies computed from a corpus of 4,500,000 words.
- [Henmon, 1924] : *French Word Book*
  $\rightarrow$ These lists were defined from frequencies (based on the word frequency effect) in the general language.
- Several issues are inherent to this approach :
  - frequency estimation is not always robust ([Thorndike, 1921] : second half of the list less robust)
  - [Michéa, 1953] highlighted that some common words in language (available words) are not well estimated.
  - Not obvious how to transform frequencies into educational levels.

Frequency lists are not really educationally-graded ressources !

Graded lists

- Graded list for L1 French is Manulex [Lètè et al., 2004] :
  - About 23,900 lemmas whose distributions have been estimated on primary schoolbooks.
  - The corpus includes 54 textbooks from CP (6 years) to CM2 (11 years)
  - Three levels were defined : CP is 1 ; CE1 is 2 and 3 spans from CE2 to CM2.

Word	Pos	Level 1	Level 2	Level 3
pomme	N	724	306	224
vieillard	N	-	13	68
patriarche	N	-	-	1
cambrioleur	N	2	-	33
Total		31%	21%	48%

Learning references

- A current reference for L2 learning is the CEFR referentials [Beacco and Porquier, 2007]
- They give more precisions than the CEFR about the specific lexical skills to learn, but...
- No distinctions are made between words within a level
- The format is not suitable for NLP approaches
- Concerns has been raised as regards the validity of these referentials (e.g. VALILEX, KELLY)

What did we learn ?

- It is acknowledged that it is possible to relate a word difficulty with some of its characteristics
- Current approaches generally focus on one or a few characteristics
  $\rightarrow$ ReSyf
- No graded resource (such as Manulex) for L2 context
  $\rightarrow$ FLELex

Collaborators

Nuria Gala, Cédrick Fairon, Patrick Watrin, Delphine Bernhard, Anaïs Tack, Laetitia Brouwers and Hubert Naets

Plan

Objectives of the FLELex project

- Offer a lexical resource describing the distribution of French words in FFL textbooks.
  $\rightarrow$ Textbooks using the CEFR scale, we get a distribution of words across the 6 levels of the CEFR.
- This distribution is learned from a corpus and the frequencies are adapted for a better estimation.
- Possible uses :
  - Targetted vocabulary learning (which word to learn at which level)
  - Comparing the frequency of usage of synonyms
  - Using it within a language model for various iCALL tasks (readability, etc.)
  - Apply it for automatic text simplification (ATS)

Methodology

1. Collect a corpus of texts from FFL textbooks
2. Tag the corpus to desambiguate forms as regards part-of-speeches
3. Compute normalized frequencies, with an adequate estimator
4. Exploring the resource

The training corpus

We collected 28 textbooks and 29 simplified books, amounting to a total of 2,071 texts and 777,000 words

Genre	A1	A2	B1	B2
Dialogue	153 (23,276)	72 (17,990)	39 (11,140)	5 (1,698)
E-mail, mail	41 (4,547)	24 (2,868)	44 (11,193)	18 (4,193)
Sentences	56 (7,072)	21 (4,130)	12 (1,913)	5 (928)
Varias	31 (3,990)	36 (4,439)	23 (5,124)	14 (1,868)
Text	171 (23,707)	325 (65,690)	563 (147,603)	156 (63,014)
Readers	8 (41,018)	9 (71,563)	7 (73,011)	5 (59,051)
Total	460 (103,610)	487 (166,680)	688 (249,984)	203 (130,752)

Genre	C1	C2	Total
Dialogue	/	/	269 (54,104)
E-mail, mail	8 (2,144)	1 (398)	136 (25,343)
Sentences	/	/	94 (14,043)
Varias	1 (272)	/	105 (15,693)
Text	175 (89,911)	48 (34,084)	1,438 (424,009)
Readers	/	/	29 (244,643)
Total	184 (92,327)	49 (34,482)	2,071 (777,835)

The tagging process

- **Goal :** obtain the lemma of every form observed in the corpus and disambiguate homographic forms with different P.O.S.
  → Using inflecting forms would imply splitting frequency density across several forms.
  → It would also imply that we consider learners unable to relate inflected forms.

- **Problem :** The tagger precision matters, otherwise we can get :
  - entries with wrong part-of-speech tag (e.g. *adoptez* PREP or *tu* ADV) ;
  - entries with a non attested lemma (e.g. *faire partir* instead of *faire partie*) ;
  - likely tags that but are erroneous in the specific context of the word.

Tagging and MWE

- One well-known limitation of taggers is their ability to extract multi-word expression (MWE) !
- MWEs includes a set of heterogeneous linguistic objects (collocations, compound words, idioms, etc.)
- Learner's knowledge of MWE lags far behind their general vocabulary knowledge [Bahns and Eldaw, 1993]
  $\rightarrow$ Therefore, including such linguistic forms in a graded-lexicon for FFL purposes appears as crucial !

The selected taggers

We selected two taggers and compared their performance :

TreeTagger

- Treetagger [Schmid, 1994] is widely used and acknowledged
- Easy to use (wrappers exists for various programming languages)
- Not anymore state-of-the-art performance and cannot detect MWEs

a CRF-based tagger

- CRF-taggers are state-of-the-art and can be trained to detect MWEs
- We used one drawing from the work of [Constant and Sigogne, 2011] and developed by EarlyTracks.

Assessing the taggers

We compared the performance of the two taggers on the same data

- Test set = 100 sentences sampled from the texts, divided in 2 batches
- Each batch was assessed by two experts, for each tagger
- Annotation of the errors was as follows :

0 no mistake ;
1 lemma is correct, but not the part-of-speech ;
2 POS-tag is correct, but not the lemma ;
3 both the POS-tag and the lemma are wrong ;
4 segmentation error (only for the CRF tagger)

Results

	TreeTagger	CRF-Tagger
correct	94.2%	95.8%
POS errors	2.6%	1%
Lemma errors	1.3%	0.5%
POS + lemma	1.9%	1.1%
Segmentation	/	1.6%

- Good agreement in general : *kappa* varied between 0.66 and 0.90
- CRF-tagger performs better than the TreeTagger !
- A few mistakes in both cases : it will produce a slight loss of probability mass !

Computing the distributions

- We used the dispersion index [Carroll et al., 1971]

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

K = CEFR level ; I = number of textbooks in level K ;
p_i = word probability in textbook i.

- Then, raw frequencies are normalized as follows :

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

where N_k = number of tokens at level k ;
$f_{min} = \frac{1}{N} \sum f_i s_i$ with f_i = word frequency in textbook i and s_i = number of words in textbook i

The two FLELex

We got two different versions of FLELex

FLELex-TT

- Includes 14,236 entries, but no MWEs !
- It is based on Treetagger and is easy to use for NLP purposes
- It has been manually checked

a CRF-based tagger

- Includes 17,871 entries, among which several thousands of MWEs
- Better performance means better estimations of frequency distributions, but segmentation errors yields to a few odd entries
- Not manually cleaned (so far)

Example of entries

lemma	tag	A1	A2	B1	B2	C1	C2	total
voiture (1)	NOM	633.3	598.5	482.7	202.7	271.9	25.9	461.5
abandonner (2)	VER	35.5	62.3	104.8	79.8	73.6	28.5	78.2
justice (3)	NOM	3.9	17.3	79.1	13.2	106.3	72.9	48.1
kilo (4)	NOM	40.3	29.9	10.2	0	1.6	0	19.8
logique (5)	NOM	0	0	6.8	18.6	36.3	9.6	9.9
en bas (6)	ADV	34.9	28.5	13	32.8	1.6	0	24
en clair (7)	ADV	0	0	0	0	8.2	19.5	1.2
sous réserve de (8)	PREP	0	0	0.361	0	0	0	0.03

The resource is freely available at

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

A few figures about FLELex

- A majority of the words are nouns in both lists (respectively 51% and 55%)
- TT-version includes 33% of hapaxes while only 26% of the entries have 10 occurrences or more.
- CRF-version includes 20% of hapaxes while 31% of the entries have 10 occurrences or more.
- We compared FLELex-TT with another lexicon : Lexique 3 [New et al., 2004]
  $\rightarrow$ Only 622 entries of FLELex-TT were missing from Lexique 3
- Correlation between total frequencies in FLELex-TT and Lexique3 is high : $0,84$

Démonstration

Perspectives

- Manually clean the CRF version
- Add a tab to the web site that would allow to directly analyze a text
- Use FLELex to predict the known/unknown vocabulary of a given reader
- Offer “FLELex” versions for other languages (currently perspectives for Swedish and Spanish)
  $\rightarrow$ What about Dutch ?
- Develop a filter to go from TreeTagger tagset towards the DELAF one (used for the CRF-tagger)

Plan

Objectives of the ReSyf project

- **Scope :** Investigate word difficulty from a NLP perspective
  $\rightarrow$ Is it possible to model the complexity of words based on their characteristics ?
- **Goals :**
  - Identify variables (predictors) that characterize 'simple' words
  - Draw from data on parkinsonian patients (language and speech impairments)
  - Develop a word difficulty model and create a graded resource of synonyms (ReSyf)
- **Use :** Generalize the difficulty levels from resources such as Manulex or FLELex to a larger vocabulary
  Integrates such model to an ATS system for vocabulary simplification

Methodology

1. Collect findings about features characterizing words :
   - Psycholinguistic studies (see below)
   - 'Simple' language (people with speech impairments)
2. Define a gold standard (a list of words with levels of difficulty)
3. Identified features are seen as a predictors of word difficulty and combined within a statistical model
4. Synonyms are graded with the model

The gold standards

- We need list of words with levels of difficulty... i.e. Manulex (L1) and FLELex (L2)
- **Problem :** Both resource defines a lexical distribution for word ; they do not associate a single level to each word !
  $\rightarrow$ Three approaches :
  - First occurrence level
  - Considering each distribution as a statistical serie and taking its first quartile
  - Same, with its mean
- The best approach seems to be the first ! ! !

Looking for variables : the Parkinsonian corpus

An analysis of 'simple' language : a 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 NAAV)
- Average length : 6.3 letters, 4.7 phonemes, 1.96 syllables
- Comparison with Lexique3 : Average length : 8.6 letters, 6.8 phonemes, 2.89 syllables
- Distribution on words in Manulex :

	Level 1	**Level 2**	**Level 3**
Total in Pk_corpus	94.3%	1.45%	1.63%

Intra-lexical variables
Predictors (1/3)

- Number of letters, phonemes, syllables
- Syllable structure (more frequent structures in Pk_corpus : V, CVC, CV, CYV)
- Consistency of sound-script relationship :
  - 0 = transparency : 'abruti' `[abRyti]`
  - $<$ 2 characters : 'abrite*r*' `[abRite]`
  - $>$ 2 characters : 'abs*ent*' `[aps@]`
- Spelling patterns (double letters, digraphs)

Intra-lexical variables
Predictors (2/3)

Various variables based on an unsupervised morphological analysis [Bernhard, 2010] :

- number of morphemes
- presence of prefixation or suffixation
- is a compound
- frequency of pref./suf.
- size of the morphological family

Example

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

Psycholinguistic variables
Predictors (3/3)

- Orthographic neighbours
- Logarithm of lexical frequencies
- Presence/absence of words in Gougenheim list
- Measures of polysemy with BabelNet [Navigli and Ponzetto, 2010] (number of synsets)
- Measures of polysemy with JeuxDeMots (yes/non) [Lafourcade, 2007] (yes/non)

Towards a difficulty model

The efficiency of each variable is first assessed in isolation, with Spearman correlation (on Manulex and FleLex) :

Id	Variables	Manulex (ρ)	FLELex (ρ)
17	Frequencies from Lexique3	-0,51	-0.53
18	% of absents from Goug. (5000)	-0,41	-0.46
18	% of absents from Goug. (4000)	-0,41	-0.47
02	Number of phonemes	0,30	0,27
15	Polysemy (JdM)	-0,29	-0.38
01	Number of letters	0,27	0,25
03	Number of syllables	0,27	0,26
4a	Number of orthographic neighbours	-0,25	-0,23
4b	Cum. Freq. of neighbours	-0,25	-0,23
16	Synset BabelNet	-0,20	-0,19
6b	Nasal voyel	0,08	0,07
14	Size of family (morphoclust_10)	-0,08	-0,05
11	Prefixation (seg_10)	0.07	0,06
08	Number of morphemes (seg_10)	0,06	0,08
06	Spelling patterns (a-d)	0,05	0,06
10	Mean freq. of suffixes (freq_seg_10)	-0,05	0,02

The model

- SVM algorithm trained on the best predictors of each feature family. (Manulex = 26 var. and FLELex = 24 var.)
- Baseline 1 : majority class in data
- Baseline 2 : frequency-based model

Liste	Model	Cost	Acc.	st. dev.
Manulex	Majority class	/	48%	/
	Freq. baseline	0,1	61%	0,4%
	Model	0,5	63%	0,7%
FLELex	Majority class	/	28,8%	/
	Freq. baseline.	0,5	39%	0,8%
	Model	0,001	43%	0,5%

Discussion on the model

- This approach combines a large amount of psychologically-grounded variables to “explain” word difficulty.
- Performance of the model are not satisfactory for educational purposes, although it beats the baseline.
- This confirms that predicting lexicon difficulty is a harder task than text readability (see Bormuth)
- Some noise in Manulex : *pomme* vs *cambrioleur*, both considered as level 1.

ReSyf : a first graded list of synonyms

- Look for synonyms in JeuxDeMots [Lafourcade, 2007] : a semantic network built by crowdsourcing (game with a purpose)
- 163,543 words and expressions with semantic or thematic relations
- Given the Manulex list (19,037 words), extraction of all the words having a synonym in JdM
- Result : 17,870 graded words with 12,687 graded synonyms

	Level 1	**Level 2**	**Level 3**
Total Manulex words in JdM	30.1%	21%	48.9%

ReSyf : a first graded list of synonyms

Example

piétiner(2) = marcher(1), fouler(3), piaffer(3), trépigner(3)

- Applying our model to JdM words absent from the Manulex graded list
- Available soon at `http://uclouvain.resyf.be`

Discussion and perspectives

policier(1) = poulet(1), flic(2), commissaire(3)
glacial(2) = froid(1), sec(1), impassible(3), imperturbable(3), insensible(3), glacé(1), polaire(2), rigoureux(2), inhospitalier(3)
yéti(3) = abominable(2) homme(1) neige(1)

- Language registers > classify levels (slang, current, formal)
- Polysemy > word sense disambiguation
- Compositionnality > evaluate the impact of opacity

References I


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


Bahns, J. and Eldaw, M. (1993).
Should We Teach EFL Students Collocations ?
System, 21(1) :101–14.


Beacco, J.-C. and Porquier, R. (2007).
Niveau A1 pour le français : utilisateur-apprenant élémentaire.
Didier.


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

References II


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., Davies, P., and Richman, B. (1971).
The American Heritage word frequency book.
Houghton Mifflin Boston.


Chall, J. and Dale, E. (1995).
Readability Revisited : The New Dale-Chall Readability Formula.
Brookline Books, Cambridge.


Constant, M. and Sigogne, A. (2011).
Mwu-aware part-of-speech tagging with a crf model and lexical resources.
In *Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World*, pages 49–56.

References III


De Coster, I., Baidak, N., Motiejunaite, A., and Noorani, S. (2011).
Teaching reading in europe : Contexts, policies and practices.
Technical report, Education, Audiovisual and Culture Executive Agency, European Commission.


Gerhand, S. and Barry, C. (1999).
Age of acquisition, word frequency, and the role of phonology in the lexical decision task.
Memory & Cognition, 27(4) :592–602.


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.


Henmon, V. (1924).
A French word book based on a count of 400,000 running words.
Bureau of Educational Research, University of Wisconsin, Madison.

References IV

Hu, M. and Nation, P. (2000).
Unknown vocabulary density and reading comprehension.
Reading in a foreign language, 13(1) :403–30.

Lafourcade, M. (2007).
Making people play for lexical acquisition with the jeuxdemots prototype.
In *SNLP'07 : 7th international symposium on natural language processing*.

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

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

References V


Michéa, R. (1953).
Mots fréquents et mots disponibles. un aspect nouveau de la statistique du langage.
Les langues modernes, 47(4) :338–344.


Millis, M. and Button, S. (1989).
The effect of polysemy on lexical decision time : Now you see it, now you don't.
Memory & Cognition, 17(2) :141–147.


Navigli, R. and Ponzetto, S. P. (2010).
Babelnet : Building a very large multilingual semantic network.
In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.


New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004).
Lexique 2 : A new French lexical database.
Behavior Research Methods, Instruments, & Computers, 36(3) :516.

References VI


Patel, V., Branch, T., and Arocha, J. (2002).
Errors in interpreting quantities as procedures : The case of pharmaceutical labels.
International journal of medical informatics, 65(3) :193–211.


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.


Richard, J., Barcenilla, J., Brie, B., Charmet, E., Clement, E., and Reynard, P. (1993).
Le traitement de documents administratifs par des populations de bas niveau de formation.
Le Travail Humain, 56(4) :345–367.


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.

References VII


Schwanenflugel, P., Harnishfeger, K., and Stowe, R. (1988).
Context availability and lexical decisions for abstract and concrete words* 1.
Journal of Memory and Language, 27(5) :499–520.


Thorndike, E. (1921).
Word knowledge in the elementary school.
The Teachers College Record, 22(4) :334–370.