Dmesure et FLELex: deux approches de la difficulté textuelle pour le français langue étrangère

Thomas François

UCL Université catholique de Louvain

Seminar in Fukuoka University

January 7th, 2015

Plan

Plan

Context

- The sector of foreign language teaching is growing and changing :
  - There is a will to optimize the costs of education while improving its quality
  - The number of professionals is insufficient relative to demand.
  - Learners want more flexibility in teaching methods (timetable, place...).

CALL has been viewed as a solution to these issues through the development of various pieces of software.

CALL and NLP

Natural language processing aims to automatically process natural language (translation, information retrieval, etc.).

It can also be used for within CALL framework for L2 teaching (and L2 learners directly) in various ways :

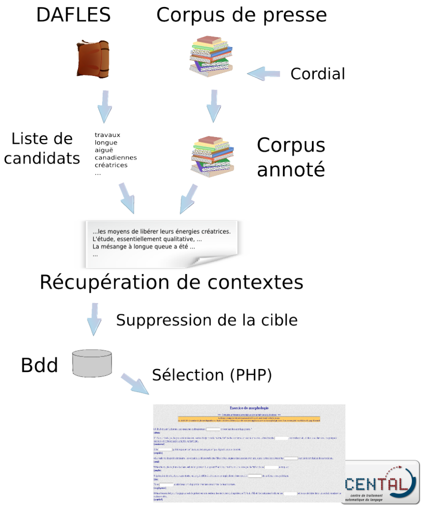
- Integration of NLP-tools within CALL software for better adaptability, intelligent feedback, or incremental content collection ;
- Automatic selection of reading materials at their level (readability) ;
- Automatic generation of reading or language exercices ;
- Automatic text simplification (ATS)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

Generation of exercises : an example

- **ALFALEX**

  [Selva, 2002, Verlinde et al., 2003]

  - Automated design of exercises on morphology, gender, collocations...
  - Difficulty of the task : 2 levels
  - Difficulty of the context is not controlled !
    It depends on the level of the corpus used.
  - http ://www.kuleuven.be/alfalex/



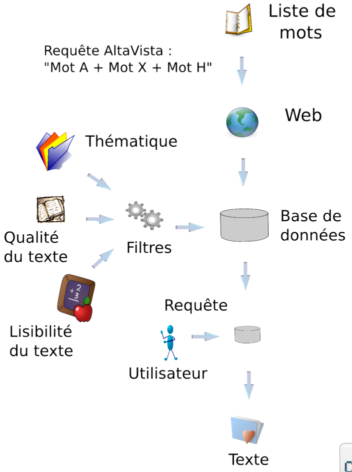
Retrieval of web texts : an example for EFL

- **REAP**

  [Heilman et al., 2008b,

  Collins-Thompson and Callan, 2004]

  - REAding-specific Practice aims at improving reading comprehension abilities through practice.
  - It integrates a SVM thematic classifier
  - Difficulty is checked using the readability formulas described in [Collins-Thompson and Callan, 2005, Heilman et al., 2008a]
  - http ://reap.cs.cmu.edu/



Readability : an example

Grammar-based Reading Difficulty Prediction

Grade level predicted: 12.0

Accuracy generally improves with text length. The software will provide estimates for texts of any length, but a minimum length of 30 words is recommended. Also, the system is generally more accurate for grade levels above 2.

Type or paste your text into the box below and press "Submit" to obtain an estimate of the difficulty of your text.

A narrow grave-yard in the heart of a bustling, indifferent city, seen from the windows of a gloomy-looking inn, is at no time an object of enlivening suggestion; and the spectacle is not at its best when the mouldy tombstones and funereal umbrage have received the ineffectual refreshment of a dull, moist snow-fall. If, while the air is thickened by this frosty drizzle, the calendar should happen to indicate that the blessed vernal season is already six weeks old, it will be admitted that no depressing influence is absent from the scene.

Submit

An estimation of the readability of the first lines of *The Europeans* (H.James). It has been assessed by the model of [Heilman et al., 2007].

Url : http ://boston.lti.cs.cmu.edu/demos/readability/index.php

Content of the talk

In this talk, we focus on two NLP-enabled approaches of CALL, specialized for L2 reading :

DMesure : a readability formula for FFL

- DMesure is a readability model that is able to associate a text with a CEFR level (for a reading task).

FLELex : a CEFR-graded lexicon for FFL

- FLELex provides a lexicon for which the distribution of each word across CEFR levels is described.

Plan

1. Introduction
2. DMesure : a readability model for FFL
   - Introduction to readability formulas
   - Conception of DMesure
   - Results
3. FLELex : a graded lexicon for FFL
   - Introduction : theoretical background
   - The making of FLELex

Plan

What is readability ?

Origin : Readability dates back to the 20s, in the U.S. It is only after 1956 that it spread in the French-speaking community.

Objective : Aims to assess the difficulty of texts for a given population, without involving human judgements.

Method : Develop tools, namely readability formulas, which are statistical models able to predict the difficulty of a text given several text characteristics.

Most famous ones are those of [Dale and Chall, 1948] and [Flesch, 1948].

Example of a formula

Formula of [Dale and Chall, 1948, 18] :

$$X_1 = 3,6365 + 0,1579\, X_2 + 0,0496\, X_3$$

where :

- X_1 : mean grade level for a schoolchild that would be able to get at least 50% to a comprehension test on this text.
- X_2 : percentage of words not in the list of Dale (3000 words).
- X_3 : mean number of word per sentence.

The independant variables X_2 and X_3 are the **predictors** or **features**).

What are the use for readability formulas ?

Readability formula have been used for :

- Selection of materials for textbooks.
- Calibration of books for children [Kibby, 1981, Stenner, 1996].
- Used in scientific experiments to control the difficulty of textual input data.
- Controling the difficulty level of publications from various administrations (justice, army, etc..) and newspapers.
- More recently, checking the output of automatic summarization, machine translation, etc. [Antoniadis and Grusson, 1996, Aluisio et al., 2010, Kanungo and Orr, 2009].

What about readability formulas for FFL ?

Common approach for foreign language contexts : apply formula designed for natives [Cornaire, 1985]

→ Denial of the specific process of L2 reading.

This approach relies on three suspect assumptions

- the understanding of readers in the L2 is comparable to that of native speakers.
- the textual features considered in L1 formulas are relevant to L2 reading (and the only relevant ones).
- the weighting of these variables can be the same in a formula for L1 and L2.

An alternative : consider the specificities of the L2 context

Some studies took into account those specificities, described by [Koda, 2005], into readability models :

- [Tharp, 1939] positions himself against the previous approach and offers one of the first specific formulas for FLE, based on cognates.
- [Uitdenbogerd, 2005] suggests a formula that also takes into account cognates :

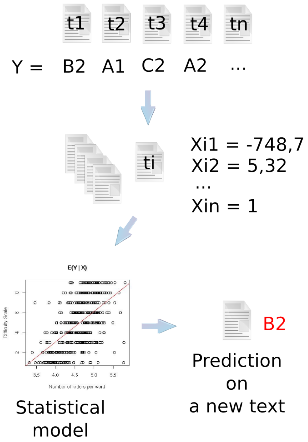
$$\mathrm{FR} = 10 * WpS - Cog$$

WpS : mean number of word per sentence.
Cog : number of cognates per 100 words.

- [Heilman et al., 2007] compare the efficiency of lexical and syntactic features in L1 and L2 context :
  $\rightarrow$ grammatical features play a more important role in a L2 model.

Conception of a formula : methodological steps

1. Collect a corpus of texts whose difficulty has been measured using a criterion such as comprehension tests or cloze tests
2. Define a list of linguistic predictors of the difficulty, such as sentence length or lexical load
3. Design a statistical model (traditionally linear regression) based on the above features and corpus
4. Validate the model



Plan

The corpus (1)

- Criterion = expert judgments = textbooks !
  → The assumption is that the level of a text can be considered the same as the level of the textbook it comes from.

- The type of criterion affects the difficulty scale used.
  → We extracted 2042 texts from 28 FFL textbooks, following the CEFR scale [Conseil de l'Europe, 2001].

The CEFR scale

It is the official EU scale for L2 education.
It has 6 levels : A1 (easier), A2, B1, B2, C1, and C2 (higher).

Corpus (2)

Not all FFL textbooks were used :

1. Have to follow the CEFR recommandations (posterior to 2001).
2. Language should be modern (arises from condition 1).
3. Intended audience : young people and adults (not children).
4. General reading : I excluded FSP textbooks.

Another selection was performed at the text level :

1. Only texts related to a reading comprehension task.
2. Instructions were not considered.

Distribution of the texts per level

	A1	A1+	A2	A2+	B1	B1+	B2	C1	C2
Activités CECR	/	/	/	/	41	39	50	63	8
Alter Ego	46	44	61	31	74	42	/	/	/
Comp. écrite	/	/	34	53	39	50	/	/	/
Connexions	34	26	/	/	/	/	/	/	/
Connexions : prep. DELF	/	11	/	12	/	/	/	/	/
Delf/Dalf	/	/	/	/	/	/	31	78	19
Festival	42	34	/	/	28	26	/	/	/
Ici	13	28	25	17	/	/	/	/	/
Panorama	31	27	50	48	56	57	41	/	/
Rond-point	3	19	4	7	21	19	76	/	/
Réussir Dalf	/	17	/	/	/	/	/	43	22
Taxi !	27	/	23	21	56	51	/	/	/
Tout va bien !	/	50	36	56	45	37	/	/	/
Total	**196**	**256**	**233**	**245**	**360**	**321**	**198**	**184**	**49**

TABLE : Number of texts per level, for each textbook series used.

Predictors from the literature

I implemented 406 variables, most of them draw inspiration from previous studies :

- **lexical** : statistics of lexical frequencies ; percentage of words not in a reference list ; N-gram models ; measures of lexical diversity ; length of the words ;
- **syntactic** : length of the sentences ; part-of-speech ratios ;
- **semantic** : personnalisation level ; idea density ; coherence level measured with LSA ;

specific to FFL : detection of dialogue.

Some of them were never experimented in a FFL (or even L2) context.

Contribution of cognitivist studies on the reading process

Psychological description of the reading process provided ideas for new predictors :

lexical : orthographic neighbors ; normalized TTR ; **number of meanings per words**.

syntactic : verbal moods and tenses ;

specific to FFL : characteristics of MWE, **acquisition steps**.

Features in bold have not been implemented so far.

Machine learning algorithms

- **Regression models** : they depend on the type of the dependant variable

Continuous	$\Rightarrow$	Linear regression
Ordinal	$\Rightarrow$	Proportional odds model (OLR)
Categorical	$\Rightarrow$	Multinomial logistic regression (MLR)

- Models based on **decision trees** :
  - Classification tree [Breiman et al., 1984]
  - Boosting [Freund and Schapire, 1996]
  - Bagging [Breiman, 1996]
- **S**upport **V**ector **M**achines [Boser et al., 1992]

Plan

Results in two steps

Our experimentation were conducted in two steps :

1. Evaluation of the predictive ability of variables used alone (= bivariate analysis).
2. Evaluation of the predictive ability of some combinations on variables (= modelisation step).

The goal : limit multicollinearity risks.

Bivariate analysis : some variables

	Test6CE			
	r	ρ	$W(p)$	$F(p)$
X75FFFDC	-0.296^{2}	-0.627^{3}	$< 0,001$	0.089
X90FFFC	-0.319^{3}	-0.641^{3}	$< 0,001$	$< 0,001$
PAGoug_2000	0.593^{3}	0.597^{3}	$< 0,001$	0.017
PA_Alterego1a	0.657^{3}	0.652^{3}	$< 0,001$	$< 0,001$
ML3	-0.56^{3}	-0.546^{3}	$< 0,001$	$< 0,001$
meanNGProb.G	0.382^{3}	0.407^{3}	0.011	0.05
NLM	0.479^{3}	0.483^{3}	0.028	0.084
NL90P	0.519^{3}	0.521^{3}	$< 0,001$	0.022
NMP	0.486^{3}	0.618^{3}	$< 0,001$	0.014
PRO.PRE	-0.181^{3}	-0.345^{3}	$< 0,001$	0.226
PPres	0.44^{3}	0.44^{3}	$< 0,001$	0.003
Pres_C	-0.355^{3}	-0.337^{3}	$< 0,001$	$< 0,001$
PP1P2	-0.408^{3}	-0.333^{3}	$< 0,001$	0.008
avLocalLsa_Lem	$0,63^{3}$	$0,63^{3}$	$< 0,001$	0,01
NACoIl	/	0.286^{3}	/	/
BINGUI	$0,462^{3}$	$0,462^{3}$	$< 0,001$	0,018

Main results from the bivariate analysis

- Each familly has at least one efficient predictor
  $\rightarrow$ idea : what if I design a formula with those variables ?
- Among those, two are traditional ones (**PA_Alterego1a** et **NMP**) and one is NLP-based (**avLocalLsa_Lem**).
- Surprisingly, some other NLP-based features are poor predictors : N-gram models (where N>1), MWE-based features, etc.
- **Specialization** : the efficiency of **PA_Alterego1a** provides a rationale for adapting readability models to specific contexts (list for FFL).

Design of the readability model

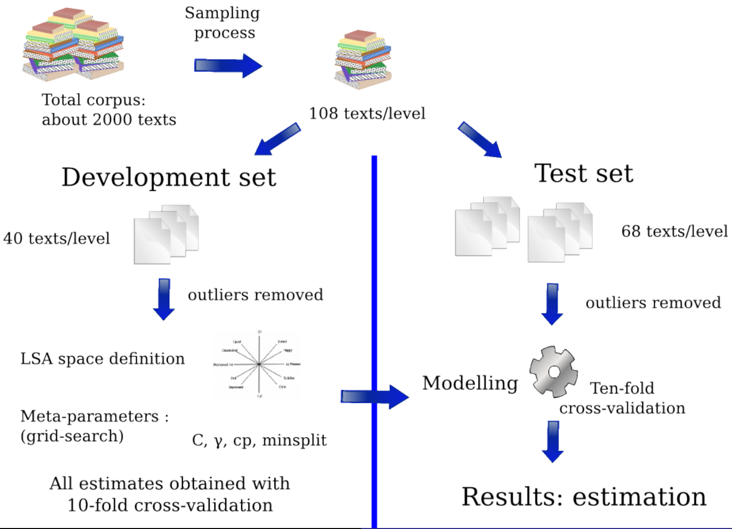
For the modelisation step, various combinations of predictors were attempted :

- Baseline (mimics classic formulas) : NMP + NLM.
- Best predictor/familly (4) : PA_Alterego1a + NMP + avLocalLsa_Lem + BINGUI.
- 2 best predictors/familly (8) : PA_Alterego1a + X90FFFC + NMP + PPres + avLocalLsa_Lem + PP1P2 + BINGUI + NAColl.

  → Assumption : maximizing the **type** of information in a minimal set.
- Automatic selection of features.
  → Assumption : maximizing the **quantity** of information.

Each set was tested with the 6 statistical algorithms.

Design of the readability model (2)



Evaluation measures

Models were evaluated with these 5 measures :

- Multiple correlation ratio (*R*).
- Accuracy (*acc*).
- Adjacent accuracy ($acc - cont$)
  → proportions of predictions that were within one level of the human-assigned level for the given text [Heilman et al., 2008a]
- Root mean square error (RMSE).
- Mean absolute error (MAE).

Main results

Model	Classifier	Parameters	*R*	*acc*	*acc − cont*	*rmse*	*mae*
Random	/	/	/	16, 6%	44, 4%	/	/
Baseline	SVM	$\gamma = 0,05; C = 25$	0, 62	34%	68, 2%	1, 51	1, 06
Model 2009	RLM	/	0, 62	41%	71%	/	/
Expert1	RLM	/	0, 70	39%	74, 2%	1, 34	0, 97
Expert2	SVM	$\gamma = 0,002; C = 75$	0, 73	41%	78%	1, 28	0, 94
Auto-OLR	OLR	/	0.71	39.6	76.1	1.33	0.96
Auto	SVM	$\gamma = 0,004; C = 5$	0, 73	49%	79, 6%	1, 27	0, 90

Best model

- $+32,4\%$ in comparison with random (*acc*) ;
- $+8\%$ in comparison with previous 2009 model (*acc*) ;
- Adjacent accuracy per level, computed on one of the 10 folds (mean is 79%)

	A1	A2	B1	B2	C1	C2
Adj. acc.	100%	71%	67%	71%	86%	83%

Contribution of the variable families

We compared models either using only one family of predictors, or including all 46 features except those of a given family :

	Family only		**All except family**	
	Acc.	Adj. acc.	Acc.	Adj. acc.
Lexical	40.5	75.6	41.1	73.5
Syntactic	39.3	69.5	43.2	78.4
Semantic	28.8	61.5	47.8	79.2
FFL	24.9	58.5	47.8	79.6

Results

- lexical and then syntactic families reach the highest performance and yield the highest loss in accuracy.
- Lexical features are the only ones to reduce the amount of critical mistakes (adj. acc.).

First conclusions

- It is the first specific formula for FFL that uses a NLP approach (and one of the few for FFL)

  $\Rightarrow$ The corpus includes a variety of text types, ensuring a wider coverage to the formula
- The criterion used (level of the textbooks according to the CEFR scale) appears questionable : the noise in the corpus can cause a poor learning.
- Our experiments suggest the (slight) superiority of SVM and logistic regression, a technique which is less demanding than the first.

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.

Plan

Plan

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 the second section of 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.
  → 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*
  → 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, Anaïs Tack

Plan

Objectives of the FLELex project

- Offer a lexical resource describing the distribution of French words in FFL textbooks.
  → 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.
  $\rightarrow$ Using inflecting forms would imply splitting frequency density across several forms.
  $\rightarrow$ 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.

Computing the distributions

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

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

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

- Then, raw frequencies are normalized as follows :

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

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

FLELex-CRF

- 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)

The end

Difficulté estimée :	A2
Votre texte :	Merci pour votre attention. Sachez que les questions et les commentaires sont les bienvenus :-)

References I

Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010).
Readability assessment for text simplification.
In *Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles.

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

Antoniadis, G. and Grusson, Y. (1996).
Modélisation et génération automatique de la lisibilité de textes.
In *ILN 96 : Informatique et Langue Naturelle*.

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

References II

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

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

Breiman, L. (1996).
Bagging predictors.
Machine learning, 24(2) :123–140.

Breiman, L., Friedman, H., Olsen, R., and Stone, J. (1984).
Classification and regression trees.
Chapman & Hall, New York.

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.

References III

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.

Collins-Thompson, K. and Callan, J. (2004).
Information retrieval for language tutoring : An overview of the REAP project.
In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 545–546.

Collins-Thompson, K. and Callan, J. (2005).
Predicting reading difficulty with statistical language models.
Journal of the American Society for Information Science and Technology, 56(13) :1448–1462.

References IV

Conseil de l'Europe (2001).
Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer.
Hatier, Paris.

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.

Cornaire, C. (1985).
La lisibilité : essai d'application de la formule courte d'Henry au français langue étrangère.
PhD thesis, Université de Montréal, Montréal.

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

References V

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

Freund, Y. and Schapire, R. (1996).
Experiments with a new boosting algorithm.
In *Machine Learning : Proceedings of the Thirteenth International Conference*, pages 148–156.

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.

References VI

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007).
Combining lexical and grammatical features to improve readability measures for first and second language texts.
In *Proceedings of NAACL HLT*, pages 460–467.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008a).
An analysis of statistical models and features for reading difficulty prediction.
In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–8.

Heilman, M., Zhao, L., Pino, J., and Eskenazi, M. (2008b).
Retrieval of reading materials for vocabulary and reading practice.
In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 80–88.

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 VII

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

Kanungo, T. and Orr, D. (2009).
Predicting the readability of short web summaries.
In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 202–211.

Kibby, M. (1981).
Test Review : The Degrees of Reading Power.
Journal of Reading, 24(5) :416–427.

Koda, K. (2005).
Insights into second language reading : A cross-linguistic approach.
Cambridge University Press, Cambridge.

References VIII

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.

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.

References IX

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.

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.

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.

Selva, T. (2002).
Génération automatique d'exercices contextuels de vocabulaire.
In *Actes de TALN 2002*, pages 185–194.

Stenner, A. (1996).
Measuring reading comprehension with the lexile framework.
In *Fourth North American Conference on Adolescent/Adult Literacy*.

References X

Tharp, J. (1939).
The Measurement of Vocabulary Difficulty.
Modern Language Journal, pages 169–178.

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

Uitdenbogerd, S. (2005).
Readability of French as a foreign language and its uses.
In *Proceedings of the Australian Document Computing Symposium*, pages 19–25.

Verlinde, S., Selva, T., and Binon, J. (2003).
Alfalex : un environnement d'apprentisage du vocabulaire français en ligne, interactif et automatisé.
Romaneske, 28(1) :42–62.