Assessing language complexity for L2 readers with NLP techniques and corpora



Thomas François

Assistant Professor at Cental, UCLouvain



NLP4CALL Workshop

30 September, 2019

Plan

1. Introduction
2. Readability assessment
3. The CEFRLex project

Plan

1. Introduction
2. Readability assessment
3. The CEFRLex project
   - Introduction to the CEFRLex project
   - Avenues for future research

Why assess text difficulty automatically ?

Which sentence is the most difficult to read and why ?

Sentence 1

At the heart of the creative process, the director is responsible for interpreting the screenplay and transforming it into a film with its own distinctive style.

Sentence 2

Interestingly, emotional tears contain a lot of chemicals and hormones which we don't find in the other kinds of tear.

Why assess text difficulty automatically ?

Which sentence is the most difficult to read and why ?

Sentence 1

At the heart of the creative process, the director is responsible for interpreting the screenplay and transforming it into a film with its own distinctive style.

From a B2 textbook

Sentence 2

Interestingly, emotional tears contain a lot of chemicals and hormones which we don't find in the other kinds of tear.

From a B1 textbook

Proficiency level assessment by humans

Traditionnally, language materials are assessed by humans, but...

- When it comes to producing a global judgment about the proficiency level or the difficulty level of a text, humans are not great about it.

Experiment 1 - non expert judges

In [François et al., 2014a], we asked to about 20 writing experts to assess the difficulty of 105 administrative texts.

Mean agreement is very low : $\alpha = 0.37$

Proficiency level assessment by humans

- When it comes to producing a global judgment about the proficiency level or the difficulty level of a text, humans are not great about it.

Experiment 2 - expert judges

In [Tack et al., 2017], we asked to 3 certified CEFR-expert Cambridge examiners to assess 299 texts produced by L2 learners of English.

Agreement is strong : $\alpha = 0.81$, but...

Proficiency level assessment by humans

Even experts sometimes disagree :

		initial level						
agreement	*%*	A1	A2	B1	B2	C1	C2	all
perfect	**43.8**	19	29	28	25	22	8	131
partial	**50.2**	17	41	27	27	26	12	150
no	**6.0**	2	2	1	4	5	4	18
		38	72	56	56	53	24	299

FIGURE – Summary of inter-examiner agreement scores

Our statistical analysis also showed that experts disagree more on the **longer texts** !

An expert that never changes its mind : the computer

There is a need for automatic assessment of difficulty for L2 learners :

- AI models use criteria that never vary (no criteria variance, but initial subjectivity).
- Various uses : information retrieval, personalised self-learning or teaching in CALL platform or classroom, helping textbook creators, etc.
- NLP allows to get information from corpora and learner data, in order to enhance the models.

Program for today

- Readability : to assess text difficulty for l2 learners
  - DMesure : an NLP-enabled readability formula for French as a foreign language (FFL)
- Difficulty identification : identifying difficult linguistic forms for L2 learners in texts
  - CEFRLex : graded lexicons to help assess lexical complexity
  - PrediComplex : level-based approach vs. personalised approach of complex word identification

Plan

1 Introduction

2 Readability assessment

3 The CEFRLex project
- Introduction to the CEFRLex project
- Avenues for future research

What is readability ?

Origin : Readability dates back to the 20s, in the U.S.

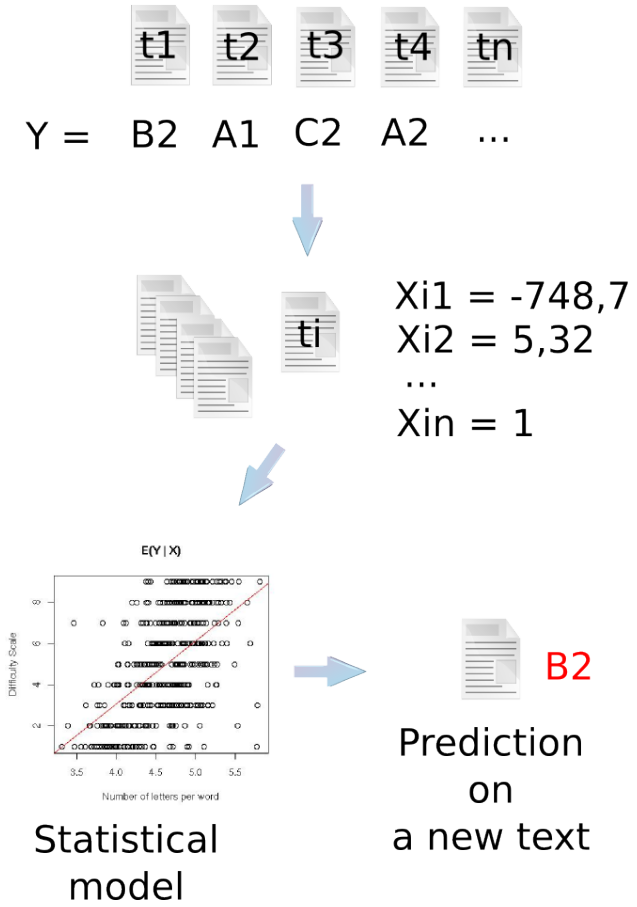
Objective : Aims to assess the difficulty of texts for a given population, without involving direct human judgements.

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

Among popular ones are those of [Dale and Chall, 1948] and [Flesch, 1948].

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 (output = single score)



Overview of performance of some AI formulas

- Performance remains unsatisfactory for commercial usage in most studies !

Étude	♯ cl.	lg.	Acc.	Adj. Acc.	R	RMSE
[Collins-Thompson and Callan, 2004]	12	E.	/	/	0.79	/
[Heilman et al., 2008]	12	E.	/	52%	0.77	2.24
[Pitler and Nenkova, 2008]	5	E.	/	/	0.78	/
[Feng et al., 2010]	4	E.	70%	/	/	/
[Kate et al., 2010]	5	E.	/	/	0.82	/
[Vajjala and Meurers, 2012]	5	E.	93.3%	/	/	0.15

- Comparison between various models in [Nelson et al., 2012] :
  - Best model from [Nelson et al., 2012] is SourceRater [Sheehan et al., 2010]

    $\longrightarrow \rho = 0.860$ on Gates-MacGinite corpus

Overview : readability formulas and deep learning

Deep learning and word embeddings have made only a timid entrance into the field

- [Cha et al., 2017] showed that bag-of-words vectors behaves poorly compared to features based on semantic-spaces.
- [Whigham et al., 2018] add a transitional network over the clustered word2vec vectors to better capture structural characteristics of texts.
- [Jiang et al., 2018] combinaison of classic features with word embeddings
  $\longrightarrow$ 95.9% accuracy for ENCT corpus (4 classes)
- [Le et al., 2018] combined word embedding with frequency embedding within a CNN.
  $\longrightarrow$ 93% accuracy for ENCT corpus (4 classes)

Our model for French as a foreign language

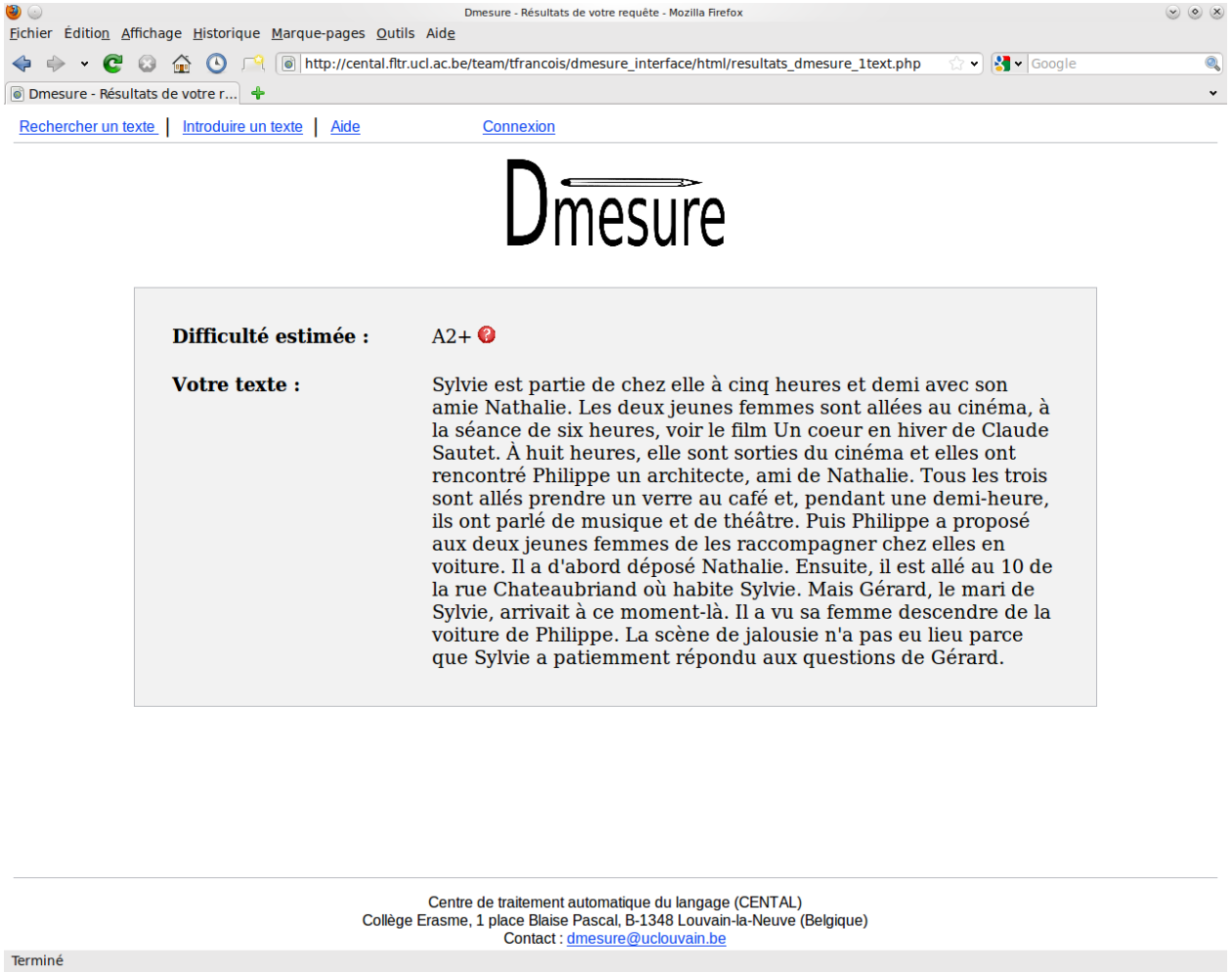
Model	Classifier	Parameters	R	acc	$acc - cont$	$rmse$	mae
Random	/	/	/	16, 6%	44, 4%	/	/
Baseline	SVM	$\gamma = 0,05$; $C = 25$	0, 62	34%	68, 2%	1, 51	1, 06
Model 2009	RLM	/	0, 62	41%	71%	/	/
Expert1	RLM	/	0, 70	39%	74, 2%	1, 34	0, 97
Expert2	SVM	$\gamma = 0,002$; $C = 75$	0, 73	41%	78%	1, 28	0, 94
Auto-OLR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

Example of output of our system



Shortcomings of readability formulas : lost information

Readability formulas produce a single score as output !

- Producing this estimate of the reading difficulty of a text is costly :
  - [Feng et al., 2010] : compute 273 features, 28 retained
  - [François and Fairon, 2012] : compute 406 features, 46 retained
  - [Vajjala and Meurers, 2012] : compute 46 features
  - ...
- Use of a lemmatiser, a tagger and/or a syntactic parser
- The results of these analyses is forgotten (ex. word frequence, complex syntactic structures, etc.)

Shortcomings of this single score : reliability

- Readability assumes that we know which texts are more difficult than other...
  $\longrightarrow$ what means “difficult”? How can we measured it?
- It is measured through another variable, easier to measure and correlated with difficulty
  $\longrightarrow$ we call it the criterion!
- Several criteria have been used in readability ... none are perfect!
  $\longrightarrow$ Human judgments, **textbooks**, comprehension tests, cloze test, reading speed, eye-tracking data, etc.

Plan

1 Introduction

2 Readability assessment

3 The CEFRLex project
- Introduction to the CEFRLex project
- Avenues for future research

Can humans better assess local difficulties ?

There are various forms of local difficulties, but we focus on word difficulty in this talk.

Which word is the most “difficult” for a L2 learner ?

- hardship vs. complication
- bin vs. basket
- nonetheless vs. nevertheless

Limits of current difficulty inventories

Level of those words according to the English Vocabulary Profile

- hardship (C1) vs. complication (C1)
- bin (B1) vs. basket (B1)
- nonetheless (C1) vs. nevertheless (B2)

Not very much informative

Common approaches to identify complex word

1. Most common approach : use **frequency lists** built on large L1 corpora.
2. **Pedagogical word lists** built by experts.
3. Studies in vocabulary learning correlating words characteristics with ease of learning [Laufer, 1997]
4. Classifying or ranking complex words with **machine learning methods**
   $\longrightarrow$ [Jauhar and Specia, 2012, Shardlow, 2013, Paetzold and Specia, 2015, Paetzold and Specia, 2016, Yimam et al., 2018]

Frequency lists : example

Text Inspector : connects words with CEFR level (EVP)



Frequency lists : limitations

Lists defined from frequencies in the general language (L1).

Are L1 frequencies suitable for all L2 contexts ?

- L1 frequencies = desirable horizon for L2 learners, BUT are they suitable for every interlanguage states ?
  $\longrightarrow$ E.g. in [Leech et al., 2001] : development (51e noun) vs. t-shirt (2999e)
- **Prevalence** : [Michéa, 1953] highlighted the existence of “available” words.

Transforming frequencies into grades

- Transforming L1 frequency lists into pedagogical ressources (with grade levels)
  $\longrightarrow$ Very simple approach : defining slices rather than setting thresholds from learner data !

Pedagogical references

- Current references for L2 learning are the CEFR reference level descriptions :
  - French : [Beacco and Porquier, 2007]
  - English : English Vocabulary profile [Capel, 2010, Capel, 2012]
  - Spanish : Plan Curricular del Instituto Cervantes
- More precise 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 [Hulstijn, 2007]

Plan

1 Introduction

2 Readability assessment

3 The CEFRLex project
- Introduction to the CEFRLex project
- Avenues for future research

Objectives of the CEFRLex project



- To offer lexical resources describing word distributions in textbooks across the 6 CEFR levels.
- Possible uses :
  - Targeted 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)

The CEFRLex project : current projects

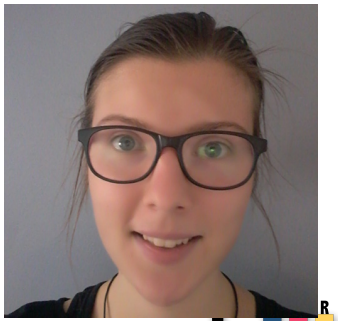
FLELex (French L2)

- Available at `http://cental.uclouvain.be/flelex/`
- Publication : [François et al., 2014b]
- Team : Thomas François, Núria Gala, Anaïs Tack, Patrick Watrin, Cédrick Fairon

EFLLex (English L2)

- Available at `http://cental.uclouvain.be/cefrlex/`
- Publication : [Dürlich and François, 2018]
- Team : Thomas François, Luise Dürlich





The CEFRLex project : current projects

SVALex (Swedish L2 - reception)

- Available at `http://cental.uclouvain.be/svalex/`
- Publication : [François et al., 2016]
- Team : Thomas François, Elena Volodina, Ildikó Pilán, Anaïs Tack

SweLLex (Swedish L2 - production)

- Available at `http://cental.uclouvain.be/svalex/`
- Publication : [Volodina et al., 2016]
- Team : Elena Volodina, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse and Thomas François







The CEFRLex project : current projects

NT2Lex - Dutch L2

- Available at `http://cental.uclouvain.be/cefrlex/`
- Publication : [Tack et al., 2018]
- Team : Anaïs Tack, Thomas François, Piet Desmet, Cédrick Fairon

ELELex - Spanish L2

- Not yet available (publication in preparation)
- Publication : scheduled for 2019
- Team : Thomas François, Barbara De Cock, Irwing Palacios





Common methodology

1. Collect a corpus of texts intended for L2 learners (from textbooks or simplified readers)
   $\longrightarrow$ The texts must be labelled with a CEFR level
2. Find the lemma and the part-of-speech tag of each word in the corpus
   $\longrightarrow$ Issue : what is a word ? MWE !
3. Estimate the frequency distribution of each lemma using a robust estimator
   $\longrightarrow$ dispersion index [Carroll et al., 1971] to normalize frequencies
4. Iterative process : manual postprocessing of the resource to correct NLP errors precedes a new frequency estimation step

For methodological details, check the papers !

Example : entries from EFLLex and ELELex

lemma	tag	A1	A2	B1	B2	C1	total
cat	NN	2.940	202.796	31.681	33.339	28.9847	65.019
empty	JJ	86.492	150.888	65.947	194.801	123.405	156.021
explore	VB	20.578	54.677	73.625	46.070	56.961	69.590
obviously	RB	0	11.034	2.589	68.463	36.665	30.689
tiresome	JJ	0	0	0	0.315	0.815	0.611
video	NN	2.467	0.556	34.825	23.802	13.248	18.431
write	VB	934.708	378.337	760.734	536.380	713.326	549.909
shopping centre	NN	0	5.040	2.589	0	0.815	1.946
sign up	VB	0	0.887	10.789	2.499	6.216	5.302

lemma	tag	A1	A2	B1	B2	C1	Total
incluir	VM	5.4	60.35	31.3	90.4	258.04	74.1
llamada	NCF	9.6	45.3	56	40.8	9.1	44.6
monumental	AQ0	0	0	1.3	2	0	1.14
malententido	NCM	0	0	0.4	0	0	0.04
acerca de	SP	0	0	4.9	21.4	18.3	8.2
al fin y al cabo	RG	0	0	0.4	15.5	10.2	4.1

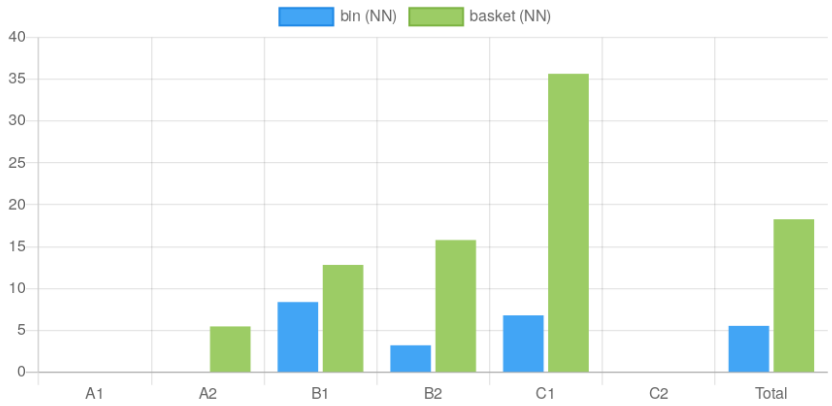
Coming back to our starting examples

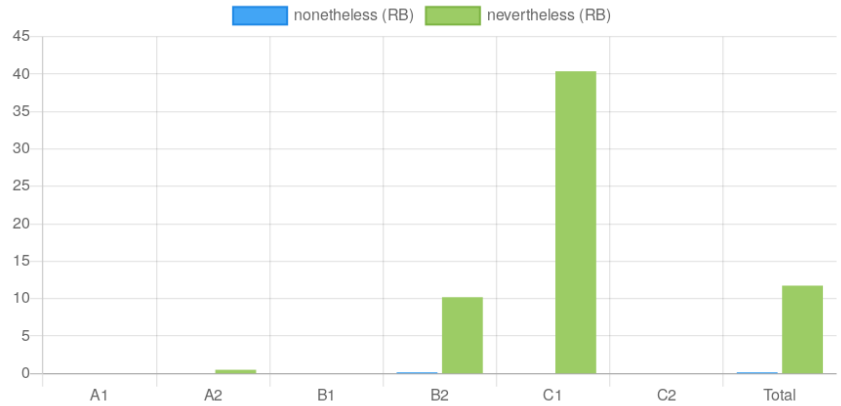
Comparing EVP and EFLLex information :

- hardship (C1) vs. complication (C1)
- bin (B1) vs. basket (B1)
- nonetheless (C1) vs. nevertheless (B2)

Not very much informative

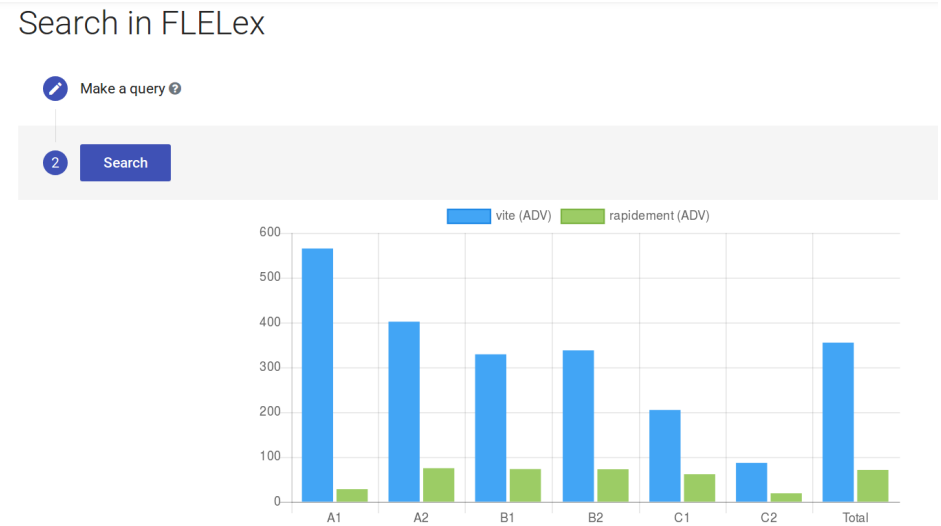
Comparing with EFLLex information





CEFRLex project : Assets

- Allow to discriminate words within a level :
- Frequency estimation from authentic pedagogical documents
- Disambiguation as regards POS-tags

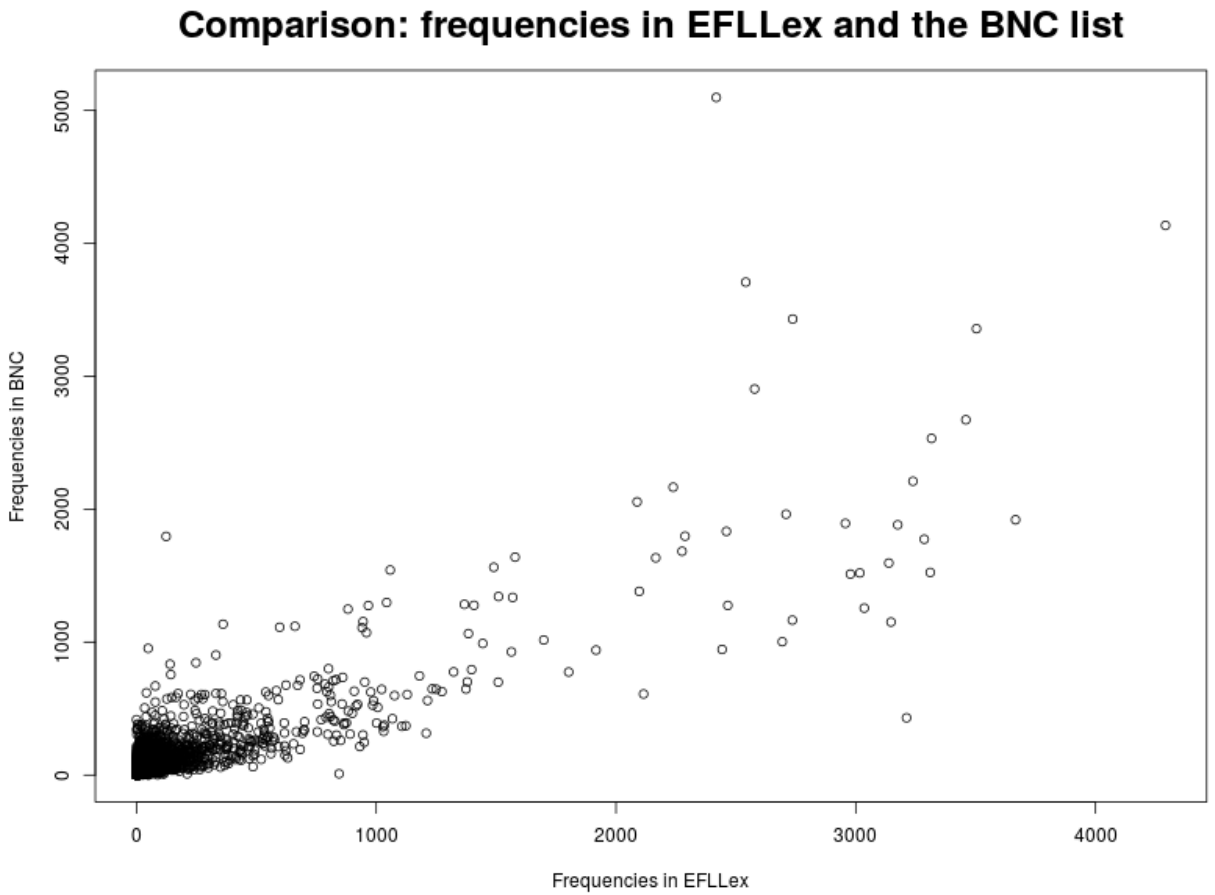


A few figures about the ressources

Resource	**# entries**	**# Hapax**	**# MWE**	***r* with list**
FLELex-TT	14,236	4,697	2,038	Lexique3 (0, 84)
EFLLex	15,280	/	/	BNC (0.97)
SVALex	15,681	/	1,450	/
ELELex	24,804	/	5,456	/
NT2Lex-CGN	15,227	/	459	Subtlex-NL (0.69)

NB : in NT2Lex-CGN, 4,431 (31%) of the single-word entries are compounds

Visualizing the correlation : BNC & EFLLex



A few figures about the ressources

Detailled figures for EFLLex per level (+ comparisons)

Level	# entries	# new entries	Hapax	>10	EVP	FLELex	SVALex
A1	2,395	2,395	893	509	601	4,976	1,157
A2	4,205	2,478	1,633	1,000	925	3,516	2,432
B1	5,607	2,740	2,366	1,003	1,429	4,970	4,332
B2	8,228	3,935	3,580	1,571	1,711	1,653	4,553
C1	9,232	3,733	4,254	1,591	N/A	2,122	3,160

Plan

1 Introduction

2 Readability assessment

3 The CEFRLex project
- Introduction to the CEFRLex project
- Avenues for future research

1. Senses instead of forms

Compare these two exemples

- I went to the bank to take some money.
- She was walking along the bank when she spotted a frog.

- Entries in most CEFRLex resources are forms (disambiguated with pos-tags)
- They cannot discriminate between these two usages, although they differ in terms of difficulty.
- Current research :
  - NT2Lex-CGN+ODWN : disambiguated version of NT2Lex [Tack et al., 2018]
  - For Swedish : [Alfter and Volodina, 2019]

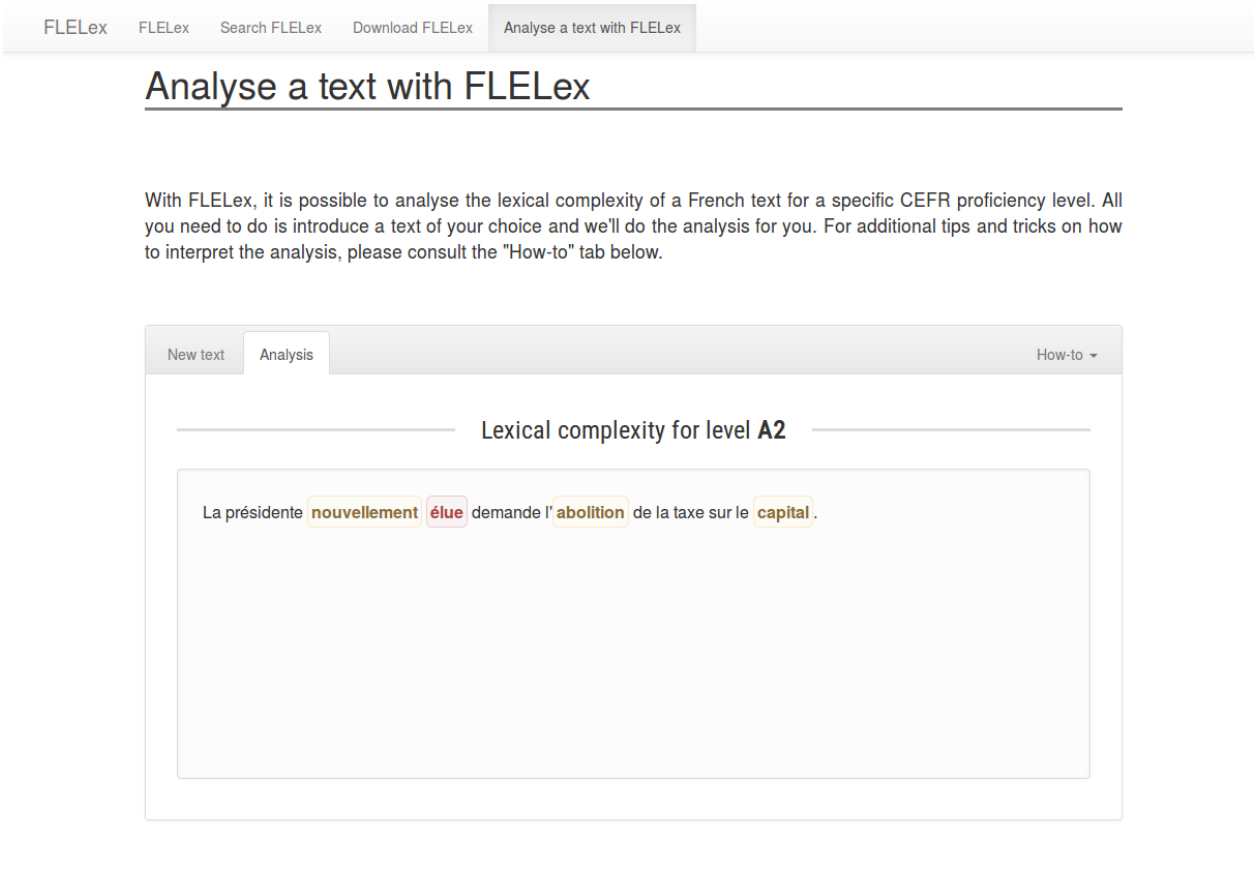
NT2Lex-CGN+ODWN

Open Dutch WordNet lists senses for lexical words (LMF format) :

<LexicalEntry id="jacht-n-1" partOfSpeech="noun">
<Sense id="r_n-18366" senseId="1" definition="hunt" synset="eng-30-00452293-n"
provenance="cdb2.2_Auto+omegawiki" annotator="">
</LexicalEntry>

<LexicalEntry id="jacht-n-2" partOfSpeech="noun">
<Sense id="r_n-18367" senseId="2" definition="yacht" synset="eng-30-04610013-n"
provenance="cdb2.2_Auto+omegawiki+wiktionary" annotator="">
</LexicalEntry>

2. Predicting unknown words in texts



Back to our starting examples

Which sentence is the most difficult to read and why ?

Sentence 1

At the heart of the creative process, the director is responsible for interpreting the screenplay and transforming it into a film with its own distinctive style.

From a B2 textbook

Sentence 2

Interestingly, emotional tears contain a lot of chemicals and hormones which we don't find in the other kinds of tear.

From a B1 textbook

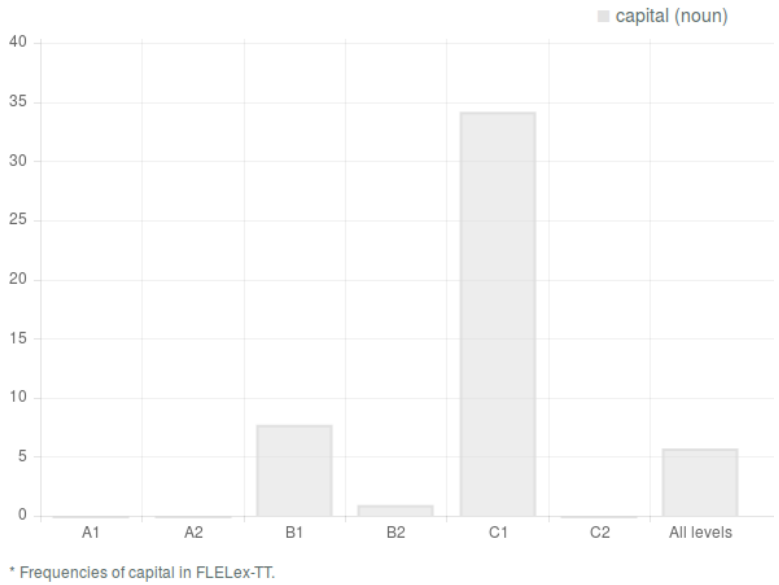
Example sentences in the text analyser



From distribution to level

Problem : How to transform a distribution into a single level ?

Example : the distribution of *capital*



* Frequencies of capital in FLELex-TT.

... is currently transformed into B1 (open issue)

Predict lexical knowledge based on FLELex

Experiment from [Tack et al., 2016a]

- Annotations collected from 4 Flemish learners of FFL (A2 and B1) on 51 texts
- Learners identify unknown words using a web interface.



Using FLELex as a way to predict lexical knowledge

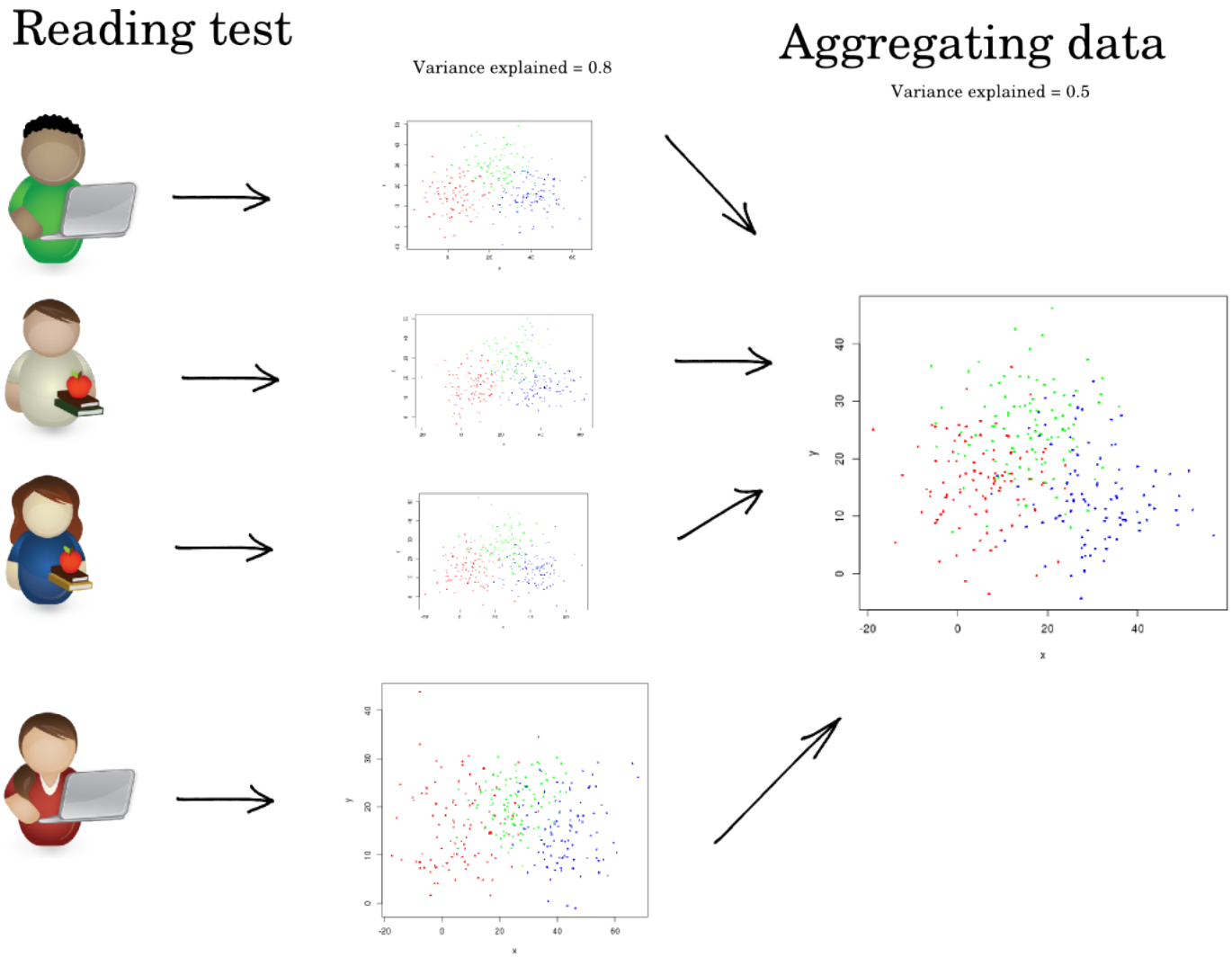
Rule to decide between known (0) and unknown (1) words :

$$c = \begin{cases} 1, & \text{if } l > L \vee \exists!\ l \\ 0, & \text{otherwise} \end{cases}$$

	Connu		Inconnu	
apprenant A2-2	95.7%	(0.92)	4.3%	(0.42)
apprenant A2-3	88.1%	(0.94)	11.9%	(0.38)
apprenant B1-4	97.0%	(0.94)	3.0%	(0.40)
apprenant B1-U	96.7%	(0.94)	3.3%	(0.37)

TABLE – Percentage of known and unknown words + recall for predictions based on FLELex.

Rationale for personalised models



Personalised models for lexical knowledge prediction (PrediComplex)

- Previous work : Alfter and Volodina (2018) ; Lee and Yeung (2018) ; Yancey and Lepage (2018), Ehara et al. (2018) ; Avdiu et al. (2019).
- [Tack et al., 2016b] (Best paper award at TALN 2016)
  - Simple features : POS, frequency distribution from FLELex, nb. letters, nb. of orthographic neighbours, spelling patterns, nb. of synsets in BabelNet (for target and neighbours)
  - Models : neural network (feedforward) trained with stochastic gradient descent

Evaluation of the models

Accuracy gain over the baseline ($\mathbf{M_E}$) (9% à 17%) :

M_P vs. M_E significant for both A2 & B1

M_L vs. M_E significant for B1

Gain of $\mathbf{M_P}$ over $\mathbf{M_L}$ not significant !

- Agreement between learners of the same level varies
- Simple features may not be discriminant enough
- Best F-1 from CW share task (2016) = 0.353 !

	A2-2			A2-3			B1-4			B1-U		
	%	F_1	κ	%	F_1	κ	%	F_1	κ	%	F_1	κ
M_P	74.6 *	0.75	0.49	74.9 *	0.72	0.50	83.0 **	0.83	0.66	83.1 **	0.82	0.66
M_L	72.4	0.74	0.45	72.2	0.71	0.44	83.8 **	0.84	0.67	81.7 **	0.80	0.63
M_E	65.7	0.54	0.31	65.3	0.52	0.31	67.6	0.55	0.34	66.3	0.51	0.32

* $p \leqslant 0.05$ ** $p \leqslant 0.01$

Conclusions

- The CEFRLex project proposes insights into the use of words by CEFR levels for different languages
- It offers various promising pedagogical uses :
  - Making finer distinctions in word use (within level or across levels)
  - Comparing the relative use of synonyms
  - Linguistics studies : association between linguistic properties and CEFR levels
  - Automatic prediction of known/unknown words of a learner.

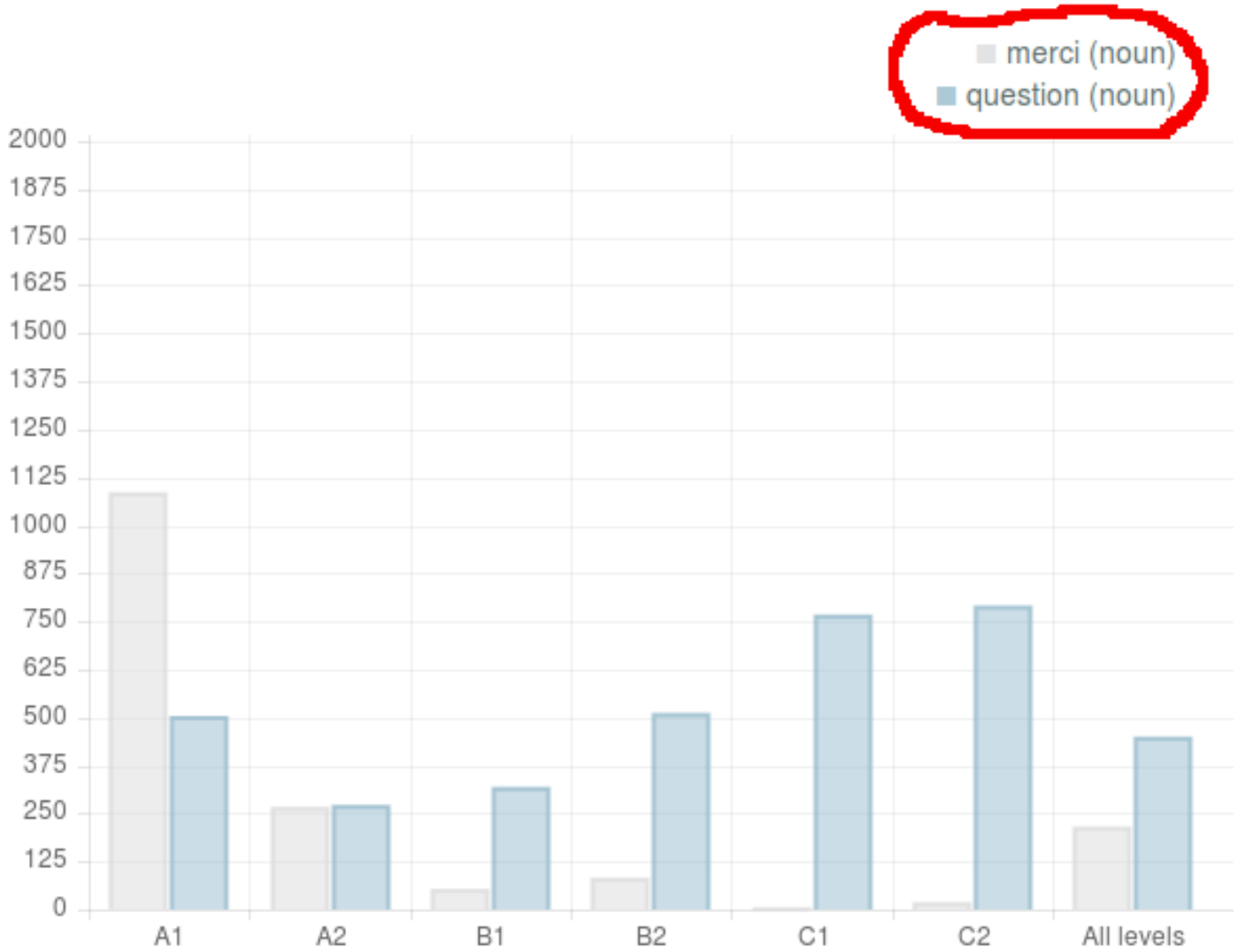
Please use CEFRLex ressources for your research and let us know your opinion/comments/results

Perspectives

- Resources for German and Portuguese in preparation (FNRS funding)
- New website still under development (recently, addition of the Spanish analyser tool)
- Validation tests with learners are scheduled
- Integrating simpler synonyms from ReSyf [Billami et al., 2018]
- Resources based on senses : available for Dutch
- A workshop scheduled for 2020, dedicated to “Building CEFR-based lexical resources”, in Louvain-la-Neuve.

Please join the project if you are interested

Thank you for your attention



Example of correct predictions

Basic words (known)

Maître Cornille ! Maître Cornille ! Ouvrez-moi la porte, s'il vous plaît.
(*Contes simplifiés*, *Le secret de maître Cornille*, apprenant B1-U)

Autrefois, il était amoureux de sa fille, mais le temps a passé et, maintenant, il l'aime comme un père.
(Contes simplifiés, Peau d'âne, apprenant B1-U)

Technical vocabulary (unknown)

L'architecture gothique dispose de nouvelles techniques : des arcs **brisés** et des **arcs-boutants**.
(*Vikidia*, *Architecture gothique*, apprenant B1-4)

Examples of incorrect predictions

Errors due to the use of lemma (and not inflected forms)

L'empereur Charlemagne **fut** enterré dans la chapelle palatine en 824.
(*Vikidia*, *Aix-la-Chapelle*, apprenant A2-3)

Son nom **savant** est l'Aquila chrysaetos".
(*Vikidia*, *Aigle royal*, apprenant A2-3)

Errors due to cognates

Les alligators se **nourrissent** d'insectes, de poissons, de **grenouilles**, de **crustacés** ou d'escargots.
(*Vikidia*, *Alligator*, apprenant A2-3)

NT2Lex-CGN+ODWN

NT2Lex-CGN+ODWN is the word-sense disambiguated version of NT2Lex [Tack et al., 2018]

- We used the list of senses from Open Dutch WordNet
- All words were disambiguated using a SVM classifier trained on the DutchSemCor (Vossen et al., 2012)
- We obtained 11,999 sense entries, of which only 1,451 entries are polysemous ...

References I


Alfter, D. and Volodina, E. (2019).
From river to bank : The importance of sense-based graded word lists.
In *EUROCALL Conference 2019*, page 178.


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


Billami, M., François, T., and Gala, N. (2018).
Resyf : a french lexicon with ranked synonyms.
In *27th International Conference on Computational Linguistics (COLING 2018)*.


Capel, A. (2010).
A1-b2 vocabulary : Insights and issues arising from the english profile wordlists project.
English Profile Journal, 1(1) :1–11.


Capel, A. (2012).
Completing the english vocabulary profile : C1 and c2 vocabulary.
English Profile Journal, 3 :1–14.

References II


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


Cha, M., Gwon, Y., and Kung, H. (2017).
Language modeling by clustering with word embeddings for text readability assessment.
In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2003–2006. ACM.


Collins-Thompson, K. and Callan, J. (2004).
A language modeling approach to predicting reading difficulty.
In *Proceedings of HLT/NAACL 2004*, pages 193–200, Boston, USA.


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

References III



Dürlich, L. and François, T. (2018).
EFLLex : A Graded Lexical Resource for Learners of English as a Foreign Language.
In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 873–879.



Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010).
A Comparison of Features for Automatic Readability Assessment.
In *COLING 2010 : Poster Volume*, pages 276–284.



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



François, T., Brouwers, L., Naets, H., and Fairon, C. (2014a).
AMesure : une formule de lisibilité pour les textes administratifs.
In *Actes de la 21e Conférence sur le Traitement automatique des Langues Naturelles (TALN 2014)*.



References IV



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)*, pages 466–477.



François, T., Gala, N., Watrin, P., and Fairon, C. (2014b).
FLELex : a graded lexical resource for French foreign learners.
In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3766–3773.



François, T., Volodina, E., Ildikó, P., and Tack, A. (2016).
SVALex : a CEFR-graded lexical resource for Swedish foreign and second language learners.
In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 213–219.



Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008).
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.



References V


Hulstijn, J. (2007).
The shaky ground beneath the cefr : Quantitative and qualitative dimensions of language proficiency.
The Modern Language Journal, 91(4) :663–667.


Jauhar, S. K. and Specia, L. (2012).
Uow-shef : Simplex–lexical simplicity ranking based on contextual and psycholinguistic features.
In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 477–481. Association for Computational Linguistics.


Jiang, Z., Gu, Q., Yin, Y., and Chen, D. (2018).
Enriching word embeddings with domain knowledge for readability assessment.
In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 366–378.

References VI


Kate, R., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R., Roukos, S., and Welty, C. (2010).
Learning to predict readability using diverse linguistic features.
In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 546–554.


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.


Le, D.-T., Nguyen, C.-T., and Wang, X. (2018).
Joint learning of frequency and word embeddings for multilingual readability assessment.
In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 103–107.

References VII


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


Nelson, J., Perfetti, C., Liben, D., and Liben, M. (2012).
Measures of text difficulty : Testing their predictive value for grade levels and student performance.
Student Achievement Partners.


Paetzold, G. and Specia, L. (2016).
Semeval 2016 task 11 : Complex word identification.
In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.


Paetzold, G. H. and Specia, L. (2015).
Lexenstein : A framework for lexical simplification.
In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 85–90.

References VIII


Pitler, E. and Nenkova, A. (2008).
Revisiting readability : A unified framework for predicting text quality.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 186–195.


Shardlow, M. (2013).
A comparison of techniques to automatically identify complex words.
In *ACL Student Research Workshop*, pages 103–109.


Sheehan, K. M., Kostin, I., Futagi, Y., and Flor, M. (2010).
Generating automated text complexity classifications that are aligned with targeted text complexity standards.
Technical report, Educational Testing Service, RR-10-28.


Tack, A., François, T., Desmet, P., and Fairon, C. (2018).
NT2Lex : A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet.
In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL 2018)*.

References IX


Tack, A., François, T., Ligozat, A.-L., and Fairon, C. (2016a).
Evaluating lexical simplification and vocabulary knowledge for learners of french : possibilities of using the flelex resource.
In *Proceedings of the Tenth conference on International Language Resources and Evaluation (LREC'16)*, pages 230–236.


Tack, A., François, T., Ligozat, A.-L., and Fairon, C. (2016b).
Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère.
In *Actes de la 23e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2016)*, pages 1–14.


Tack, A., François, T., Roekhaut, S., and Fairon, C. (2017).
Human and automated cefr-based grading of short answers.
In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179.

References X


Vajjala, S. and Meurers, D. (2012).
On improving the accuracy of readability classification using insights from second language acquisition.
In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173.


Volodina, E., Pilán, I., Llozhi, L., Degryse, B., and François, T. (2016).
Swellex : second language learners' productive vocabulary.
In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå*, number 130, pages 76–84. Linköping University Electronic Press.


Whigham, P., Chugh, M., and Dick, G. (2018).
Measuring language complexity using word embeddings.
In *Australasian Joint Conference on Artificial Intelligence*, pages 843–854. Springer.

References XI



Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Tack, A., and Zampieri, M. (2018).
A report on the complex word identification shared task 2018.
In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL 2018 Workshops)*, pages 66–78. Association for Computational Linguistics ; Stroudsburg,.