Predicting the difficulty of words for L2 readers

An empirical investigation into meaning-based and personalized word-level readability assessment with lexicon, statistical, and machine learning methods

Anaïs Tack

Profs. Drs. Cédrick Fairon, Piet Desmet, Thomas François

CENTAL, UCLouvain ITEC, imec research group at KU Leuven Aspirante F.R.S.-FNRS

Séminaire du CENTAL
April 30, 2020

Assessing the readability of words for L2 learners

- vocabulary is key to achieving successful reading
  (Jeon & Yamashita, 2014)
- ensure the readability at the word level
  1. identify/predict lexical difficulty in reading
  2. select/enhance the reading material

- **second language acquisition (SLA)**
  e.g., acquisitional complexity, ...
- **computational linguistics (NLP)**
  e.g., formalize complexity, automatic simplification
- **(intelligent) computer-assisted language learning (CALL)**
  e.g., effectiveness of educational technology

Identifying lexical difficulty in reading: An example

*Le village dans son **puits** de **rocher** n'était pas encore **noyé** sous la neige, bien qu'elle **vînt** tout près de lui, arrêtée net par les forêts de sapins qui protégeaient ses environs. Ses maisons basses ressemblaient, de là-haut, à des pavés, dans une prairie.*

— The Inn by Guy de Maupassant

Threshold methods

word length (e.g., $\geqslant 3$ syllables) (Gunning, 1952)

Lexicon methods

word frequency (e.g., basic vocabulary list) (Gougenheim et al., 1964)

Empirical & statistical methods

- observe from learner data
- engineer features, machine learning

Limitation #1: Contextualization

Lexicon methods

- same tally for word forms that are inherently polysemous (Tharp, 1939)

Empirical & statistical methods

- rankings, comparative judgments, ... on vocabulary lists (Gooding et al., 2019; Lee & Yeung, 2018; Maddela & Xu, 2018)

Meaning-based approach

1. frequency of disambiguated word senses
2. measuring and predicting difficulty of reading words in context

Limitation #2: Personalization

One-size-fits-all for optimal data collection

- splitting data among annotators (G. H. Paetzold & Specia, 2016)
- aggregated data from Amazon Mechanical Turk (Yimam et al., 2018)
- rules out variance and outliers (Dörnyei, 2009)

Personalized approach

1. learner-specific word-level readability measurements
2. link factors of lexical complexity to the learner
3. integrate learner characteristics in predictions

Structure of the thesis

Part 1: Status quaestionis

1. Predicting lexical competence in L2 reading
2. Automated identification of difficult words

Part 2: Measuring lexical difficulty

1. *A priori* knowledge: CEFR-graded word frequencies
2. *A posteriori* knowledge: Noticing difficulty while reading

Part 3: Predicting lexical difficulty

1. A mixed-effects analysis of indices of lexical complexity
2. Contextualized and personalized deep learning

Previous syntheses on vocabulary and L2 reading

L2 reading comprehension vocabulary is a strong correlate ($r = .79$)

- "a language problem rather than a reading problem" (Jeon & Yamashita, 2014, p. 196)
- skills needed to achieve a reading ease on par with natives (Melby-Lervag & Lervag, 2014)

incidental vocabulary acquisition unconscious learning of new words while engaging in meaning-focused tasks (Krashen, 1989)

1. Superior to other modes of learning? (Hulstijn, 2001; Raptis, 1997)
2. What factors contribute to vocabulary acquisition? (Huckin & Coady, 1999)
3. How effectively can we (technologically) enhance the reading input? (Abraham, 2008; Vahedi et al., 2016; Yun, 2011)

Lexical competence as a criterion variable L2 reading

Previous literature reviews have looked at either:

- the effect of vocabulary on L2 reading comprehension
- the effect of reading on L2 vocabulary acquisition

Research aims and questions

Methodological synthesis of studies statistically examining lexical competence as a criterion/dependent variable in L2 reading

1. What is the scope of publications and studies?
2. How has lexical competence been statistically modeled?
   - What measurements have been used?
   - What predictors have been tested?

A systematic scoping review

Tack, François, Desmet, and Fairon (2020, in preparation)

Inclusion/exclusion criteria

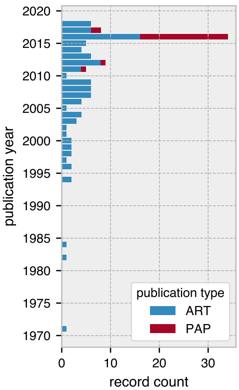
Population learners of a foreign language

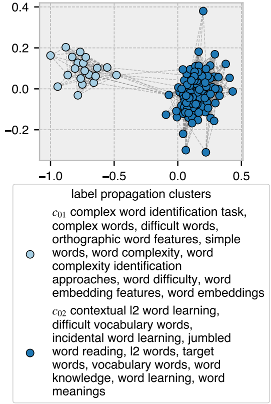
Concept
- lexical competence triggered during reading
- silent reading of words in context

Method
- empirical studies (with all reading conditions)
- dependent variable in statistical tests

1. searched in Web of Science, ProQuest, ACL Anthology
2. retrieved 2,209 records
3. selected 125 publications
4. analyzed 134 studies

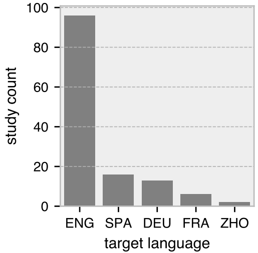
RQ1: Citation analyses of publications

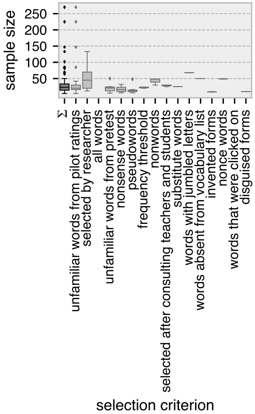




- sparsely connected bibliographic coupling network
- two research clusters
  1. applied linguistics (SLA, CALL, $c_{02} = 100$)
  2. computational linguistics (NLP, $c_{01} = 21$)
- need to bridge the gap

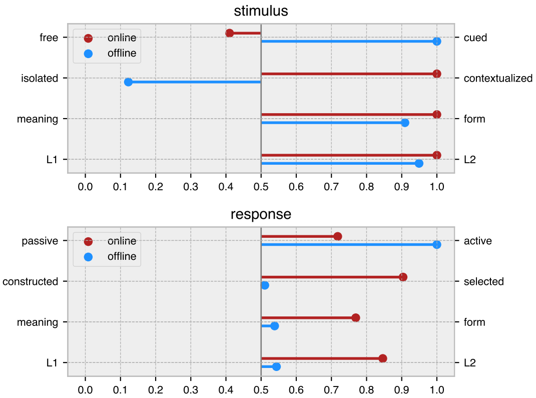
RQ1: Descriptive analyses of studies





- convenience participant samples
  - English L2
  - intermediate learners
  - university students/staff
- small samples of vocabulary
  - incidental vocabulary acquisition
  - unfamiliar (pilot/pretest)
  - non-existent words

RQ2: Measurements of lexical competence



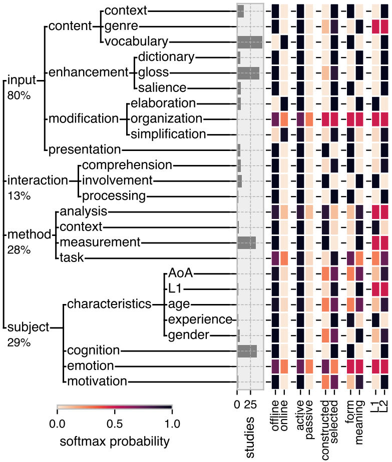
79% offline procedures (i.e., before/after reading)

- decontextualized stimuli
- form/meaning recognition/recall

21% online procedures (i.e., while reading)

- contextualized stimuli
- selected form responses

RQ2: Predictors of lexical competence



- mainly studies on effect of nature of the input
  - vocabulary traits
  - input enhancement
- fewer studies on predictors related to
  - learner
  - method of collection/analysis
  - interaction during task
- sparsity predictors and measurements

Key takeaways of the systematic review

RQ1: Scope of publications and studies

- fragmented between applied and computational linguistics
- small samples of vocabulary tested

RQ2: Lexical competence

- mainly assessed in a decontextualized manner
- few online procedures of meaning recall/recognition
- mainly input-related predictors
- not many predictors related to the learner

First CWI Shared Task (G. Paetzold & Specia, 2016)

In 1832 his family emigrated **thence** to Belleville, Ontario, where he **apprenticed** with the printer at the town newspaper, The Belleville Intelligencer.

System	Type	F_1	G
SV000gg	ensemble	0.25	0.77
PLUJAGH	threshold	0.35	0.61
CoastalCPH	neural	0.11	0.51

- data
  - Simple English Wikipedia
  - 9,200 sentences
  - distributed among 400 non-natives
  - judge comprehension difficulty
- systems
  - ensemble learning on features of lexical complexity
  - neural networks did not perform well (small training set)
  - issues with data collection

    (Zampieri et al., 2017)

Second CWI Shared Task (Yimam et al., 2018)

Both	China	and the	Philippines	flexed	their	muscles	on	Wednesday
	simple		simple	complex	complex	simple		simple
	0.0		0.0	0.4	0.25	0.0		0.0

- data
  - English, German, Spanish, French
  - L1 & L2 speakers (Amazon Mechanical Turk)
  - multi-word annotations
  - binary and probabilistic classification
- systems
  - ensemble learning with complexity features (EN)
  - neural networks top-tier performance
  - cross-lingual complexity assessment (EN/DE/ES > FR)

Is it the identification of "complex" or "difficult" words?

Complex word identification

- evolution in automatic lexical simplification
- distinguish complex and simple words
  1. based on edit histories Simple Wikipedia (Shardlow, 2013)
  2. based on user annotations (G. Paetzold & Specia, 2016; Yimam et al., 2018)

In what follows, we will define:

complexity aspects of the target language (form, meaning, function) (Kortmann & Szmrecsanyi, 2012)

difficulty for the learner

Two kinds of measures

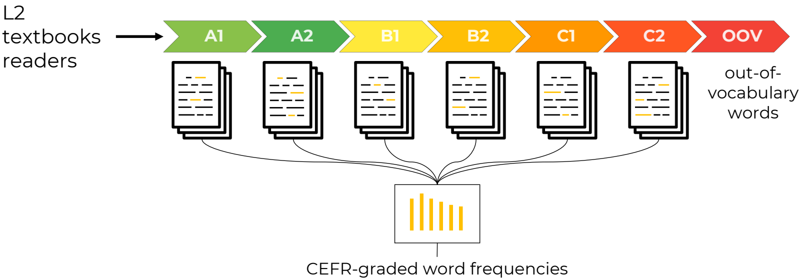
A priori knowledge: reading material in L2 textbooks and readers

- graded per CEFR (Council of Europe, 2001) level
- expert knowledge of readability for learners
- distributions of word occurrence across levels
- graded frequency lexicons

A posteriori knowledge: self-paced reading tasks

- online measurement of subjective judgment
- what triggers learners to notice (highlight) difficulty

A priori knowledge of lexical difficulty



CEFRLex project

- FLELex for French L2 (Francois et al., 2014)
- SVALex for Swedish L2 (François et al., 2016)
- ...

NT2Lex

Tack, François, Desmet, and Fairon (2018)

A lexicon with CEFR-graded frequencies for Dutch L2

- textbooks and readers, A1 to C1 levels, 461,088 tokens in total
- preprocessing
  1. part-of-speech tagging (van den Bosch et al., 2007)
  2. word-sense disambiguation on DutchSemCor
  3. linkage to Open Dutch WordNet (Postma et al., 2016)

Related to work on readability for Dutch L1

- text-to-pictograph translation, with word-sense disambiguation (Sevens et al., 2016)
- automatic lexical simplification (Bulté et al., 2018)

CEFR-graded word frequencies

NT2Lex, a graded lexicon linked to Open Dutch WordNet

The added value of word-sense disambiguation

w/o semantic disambiguation — 15,227 entries

lemma	part of speech	sense	gloss	A1	A2	B1	B2	C1
omgangstaal	n	?	'vernacular'				26	
pakken	v	?	'grab, …'	708	685	398	19	

w/ semantic disambiguation — 17,743 entries

lemma	part of speech	sense	gloss	A1	A2	B1	B2	C1
omgangstaal	n	1	'vernacular'				26	
pakken	v	1	'grab'	35	117	101	5	
⋮	⋮	⋮	⋮					
pakken	v	10	'defeat'		51	12		

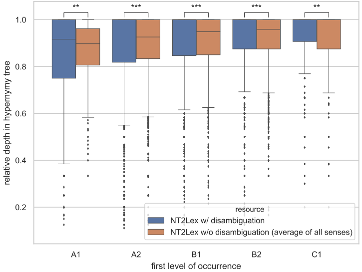
Relative depth in WordNet hypernymy tree

Index of semantic complexity

distinguish general ($d = 0$) from specific ($d = 1$) words

pakken (verb)

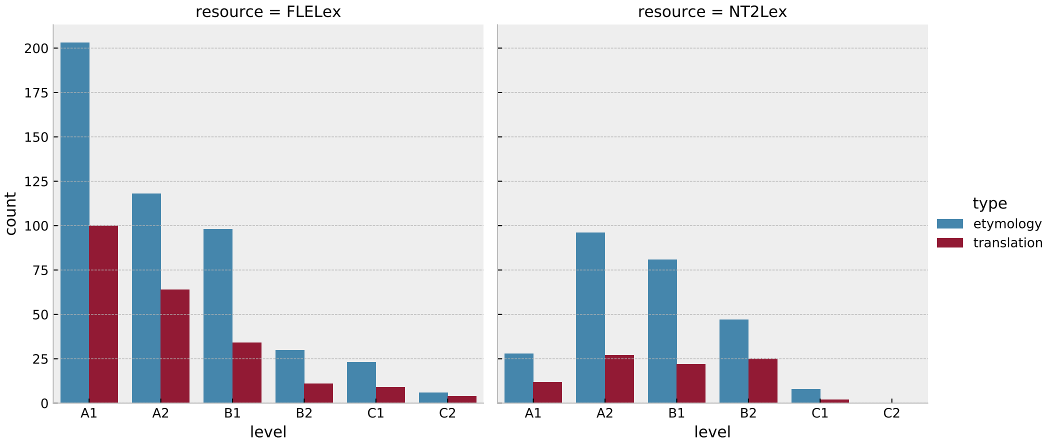
$d = .17$ get into one's hands, take physically
$d = .83$ put at a disadvantage; hinder, harm



averaged vs. disambiguated (Kruskal-Wallis tests per level)

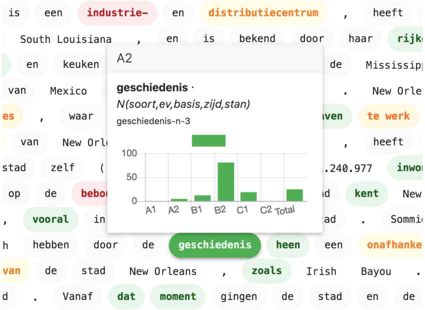
- overestimates complexity of basic words ($H_{A1} = 7.27, p = .007$)
- underestimates complexity of advanced words ($H_{C1} = 6.91, p = .009$)

Cognates in FLELex (w/ nl) and in NT2Lex (w/ fr)



Lexicon-based identification of difficult words

http://cental.uclouvain.be/nt2lex/



- enhanced lexicon-based method
  - basic lists > graded lists
  - frequency of word senses

Limitations

- one-size-fits-all approach
- no full coverage of WSD (76% of adj., adv., nouns, verbs)
- no one-to-one correspondence between resources
- further understanding needed in editorial choices

A posteriori knowledge of lexical difficulty

construct
- difficulty to decode word form
- difficulty to comprehend word meaning

behaviors
- eye gaze, movements, fixations (Štajner et al., 2017)
- misreading in read-aloud data (Gala et al., 2020)
- subjective judgment (notice difficulty)
- ...

tasks
- silent, self-paced reading of words in context
- ...

Empirical study on French L2 learners

Instructions to highlight words (≈ meaning recognition)

1. I don’t remember having seen this word before.
2. I have seen this word before, but I don’t know what it means.
3. I can’t find a synonym/explanation for this word.
4. I can’t translate this word in my native language.
5. I need to use a dictionary to understand the word.

Trial 1

- 9 subjects (A2/B1, CH/ES/JA/NL)
- 51 texts (extensive reading)

Trial 2

- 47 subjects (A2/B1/B2/C1)
- 5 texts (different per level)

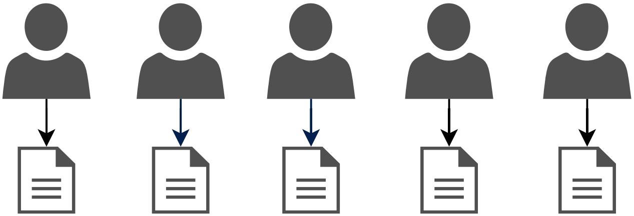
Noticing difficulty in self-paced reading

Observing difficulty as a binomial distribution

$y \sim B(N, p)$

					Subjects		
Dataset	Lang.	Pop.	N	p	M_p (SD)	$\min_p$	$\max_p$
CWI2016 (test)	EN	L2	88 221	0.047			
CWI2018 (test)	EN	L1 + L2	2095	0.383			
	EN	L1 + L2	1287	0.424			
	EN	L1 + L2	870	0.505			
	DE	L1 + L2	959	0.392			
	ES	L1 + L2	2233	0.406			
	FR	L1 + L2	2251	0.292			
Trial 1	FR	L2	189 084	0.053	0.053 ± 0.036	0.014	0.119
Trial 2	FR	L2	72 970	0.040	0.041 ± 0.032	0.010	0.151

The need to take into account random effects



	Trial 1				Trial 2			
NULL MODEL	β	*SE*	e^{β}	95% CI	β	*SE*	e^{β}	95% CI
(Intercept)	-3.07^{***}	0.17	0.05	[0.03, 0.07]	-3.37^{***}	0.20	0.03	[0.02, 0.05]
τ_{00} pro_level:sbj_id	0.45				0.34			
τ_{00} pro_level					0.14			
ICC	0.12				0.13			

*** $p < .001$

Model selection

$$\ln\left[\frac{P(Y_{ij}=1\mid u_{0j})}{P(Y_{ij}=0\mid u_{0j})}\right]=\beta_{0j}+\beta_{1j}X_{1ij}+\ldots+\beta_{mj}X_{mij}$$

$$\beta_{0j}=\gamma_{00}+\gamma_{01}Z_{1j}+\ldots+\gamma_{0n}Z_{nj}+\mathbf{u_{0j}}$$

$$\beta_{mj}=\gamma_{m0}$$

variables X features of lexical complexity
Z features related to learner (proficiency level, L1, ...)

definition u_{0j} random intercept (variability in extent of effect)

selection
- standard scaling $\sim N(0,1)$
- remove multicollinearity with VIF ($\geqslant 4$)
- stepwise forward selection with AIC

Features of lexical complexity

More than 200 predictors tested

Form
- length of the word (in characters, letters, syllables, stem)
- character *n*-grams (sequence likelihood, entropy)
- OLD20 norm (average orthographic Levenshtein distance)

Meaning
- FastText dimensions, vector similarity
- WordNet (hypernymy, hyponymy, synonymy, ...)

Use
- word *n*-gram likelihood in FRCOW16
- morphosyntactic function (part of speech, category)
- syntactic dependency function (depth, head distance, ...)
- frequency in Manulex and FLELex

Other
- Lexique3 norms
- order of occurrence, exposure, spacing in reading task
- readability thresholds (polysyllables, basic vocabulary)
- etymological information (borrowing, cognates, ...)

Generalized linear mixed-effects analysis

`lme4::glmer` with random intercepts

	Trial 1				Trial 2			
FULL MODEL	β	SE	e^{β}	95% CI	β	SE	e^{β}	95% CI
(Intercept)	-4.62^{***}	0.27	0.01	[0.01, 0.02]	-4.08^{***}	0.32	0.02	[0.01, 0.03]
ngr.word.1.ngr.frcow16ax.surprisal	0.89^{***}	0.02	2.43	[2.36, 2.51]	1.18^{***}	0.03	3.26	[3.07, 3.46]
rea.list.stopwords	-1.48^{***}	0.06	0.23	[0.20, 0.26]				
occ.expo.docu.l	-3.74^{***}	0.20	0.02	[0.02, 0.04]	-0.45^{***}	0.07	0.64	[0.56, 0.73]
res.FLELex-TT.A1_SFI	-0.38^{***}	0.02	0.69	[0.66, 0.71]	-0.77^{***}	0.04	0.46	[0.42, 0.49]
msy.categ_v	0.35^{***}	0.03	1.42	[1.34, 1.50]				
msy.categ_a	0.18^{***}	0.03	1.20	[1.12, 1.28]				
ety.borr	-0.18^{***}	0.01	0.83	[0.81, 0.86]	-0.23^{***}	0.03	0.80	[0.75, 0.85]
msy.categ_r	-0.29^{***}	0.07	0.75	[0.66, 0.85]	-1.41^{***}	0.16	0.24	[0.18, 0.33]
msy.categ_e					-3.17^{***}	0.11	0.04	[0.03, 0.05]
msy.categ_g					-1.57^{***}	0.11	0.21	[0.17, 0.26]
res.Manulex.G1_SFI					0.24^{***}	0.03	1.27	[1.19, 1.35]
τ_{00} pro_level:sbj_id	0.61				0.51			
τ_{00} pro_level					0.36			
ICC	0.16				0.21			
Marginal R^2	0.84				0.51			
Conditional R^2	0.86				0.61			

$^{***}p < .001$

Variance explained by 300-dimensional embeddings

	Trial 1 R^2	Trial 2 R^2
ngr.word.1.ngr.frcow16ax.surprisal	0.87	0.90
rea.list.stopwords	0.95 ^{T}	
occ.expo.docu.l	0.27	0.51
res.FLELex.TT.A1_SFI	0.72	0.70
msy.categ_v	0.92 ^{T}	
msy.categ_a	0.75 ^{T}	
ety.borr	0.44	0.48
msy.categ_r	0.84 ^{T}	0.94 ^{T}
msy.categ_e		0.97 ^{T}
msy.categ_g		0.98 ^{T}
res.Manulex.G1_SFI	0.71	

^{T} Tjur's R^2 on logistic regression with liblinear solver

- distributional representation
- learned hidden layers in neural network
- FastText, includes subword information (Bojanowski et al., 2017; Grave et al., 2018)
- most features are captured by these dimensions
- except for
  - incremental processing (frequency of exposure)
  - etymology (borrowing)

Introduction

Systematic literature review

Predicting lexical competence in L2 reading

Automatic identification of difficult words

Measuring lexical difficulty

CEFR-graded word frequencies

Noticing difficulty in self-paced reading

Predicting lexical difficulty

Generalized linear mixed-effects analysis

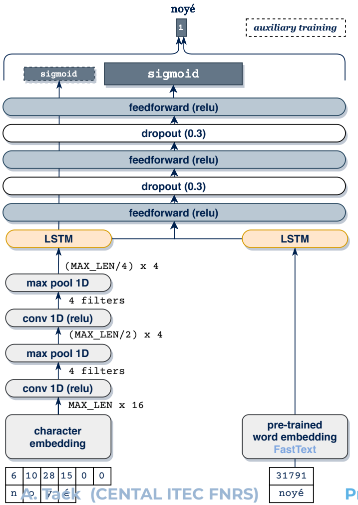
Deep learning of lexical difficulty

Conclusion

References

Appendices

Baseline Deep Neural Network



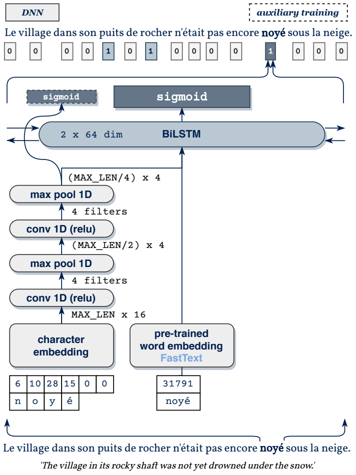
Previous work by De Hertog and Tack, 2018

- top-tier performance on EN & ES at CWI2018 shared task
- limitations
  - x recent word embeddings
  - x contextualized (sequence) learning
  - x personalized model

Adapted, distilled implementation

- TensorFlow's Keras API
- optimization on binary cross-entropy loss
  - Adam algorithm
  - balanced class weights
  - early stopping on held-out data (10%)

Two enhanced Deep Neural Networks

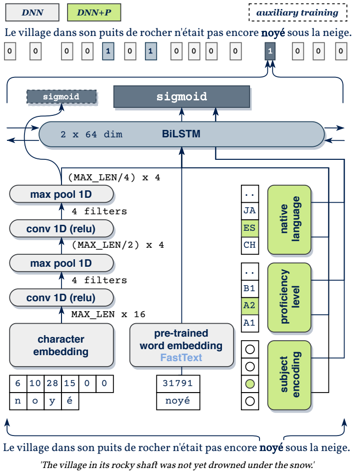


Contextualized DNN [DNN]

Bidirectional Long Short-Term Memory

- character embeddings, convolutions
- pre-trained FastText embeddings

Two enhanced Deep Neural Networks



Contextualized DNN [DNN]

Bidirectional Long Short-Term Memory

- character embeddings, convolutions
- pre-trained FastText embeddings
  (Bojanowski et al., 2017; Grave et al., 2018)

Personalized DNN [DNN+P]

+ subject encoding (ID)
+ proficiency level (CEFR)
+ native language (L1)

Performance analysis

Current SOTA benchmark metrics are sensitive to uncertainty in true distributions

$y_{\text{subject}} \sim B(N, p)$

- one-factor-at-a-time analysis on prior p (constant N, constant model $P[y = 1] = 1$)
- insensitive to uncertainty

Phi / MCC correlation coefficient

- binarization
- robust on class imbalance (Boughorbel et al., 2017)

Tjur's *D* coefficient of discrimination

- differences in mean $\hat{P}(y = 1)$
- between $y = 1$ and $y = 0$

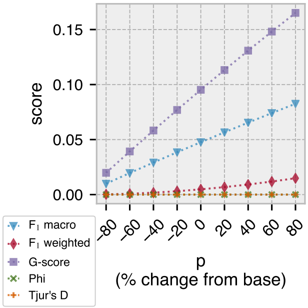


Figure: Performance after percent changes from $p_{\text{base}} = 0.05$

Stratified 10-fold cross-validation

Table: Friedman repeated measures (median)

	Ablation	ϕ	D	$\bar{\hat{P}}_1$ [a]
MODELS				
DNN+P	Full	0.38	0.67	0.85
	– L1	0.37	0.65	0.81
	– ID	0.35	0.67	0.86
	– Level	0.35	0.63	0.83
DNN	Full	0.33	0.67	0.85
	– CharCNN	0.32	0.65	0.87
	– FastText	0.19	0.45	0.88
BASELINES				
DeHertog2018		0.41	0.58	0.65
CEFR		0.12	N/A	N/A
Constant		0.00	0.00	1.00

[a] Average estimated probability of difficulty on $y = 1$

Contextualization

DNN outperforms DeHertog2018
$z = 4.78, p < .001$

Personalization

DNN+P outperforms DNN
$z = 3.95, p < .001$

Ablation worsens performance, esp.:

- word embeddings
- proficiency level
- subject ID

Key takeaways I

Focus #1: Contextualization

Literature few contextualized measurements

Measure graded lexicon with word-sense disambiguation

- better estimate of word frequency
- better estimate of semantic complexity

Prediction contextualized DNN (BiLSTM) achieves

- better overall certainty of difficulty
- better discriminative power ($D = .67$)

Key takeaways II

Focus #2: Personalization

Literature few subject- and task-related predictors

Measure how learners notice difficulty

- variability that was not previously accounted for
- larger samples of vocabulary tested

Prediction mixed models and deep learning

- substantial explanatory power with shallow features
- sensitivity of SOTA metrics to learner variability
- personalized DNN achieves better correlation ($r_\phi = .38$)

Future directions

Measuring lexical difficulty

1. enhance experimental design with task-related factors
2. contrast subjective noticing of difficulty with other online measures
3. authentic learning context (e.g., NedBox)

Predicting lexical difficulty

1. continue exploration of fine-tuning with CamemBERT
2. perceived effectiveness of predictions

Thank you!
Any questions?

References I

- Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis [WOS:000268814100001]. *Computer Assisted Language Learning*, *21*(3), 199–226. https://doi.org/10.1080/09588220802090246
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10/gfw9cs
- Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric (Q. Zou, Ed.). *PLOS ONE*, *12*(6), e0177678. https://doi.org/10/gbgsqd
- Bulté, B., Sevens, L., & Vandeghinste, V. (2018). Automating lexical simplification in Dutch. *Computational Linguistics in the Netherlands Journal*, *8*, 24–48. Retrieved November 28, 2019, from https://clinjournal.org/clinj/article/view/78

References II

- Council of Europe. (2001). *Common European Framework of Reference for Languages*. Cambridge University Press.
- De Hertog, D., & Tack, A. (2018). Deep Learning Architecture for Complex Word Identification. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 328–334. Retrieved 2018, from https://www.aclweb.org/anthology/W18-0539
- Dörnyei, Z. (2009). *The Psychology of Second Language Acquisition* [OCLC: 695695431]. Oxford Univ. Press.

References III

- Francois, T., Gala, N., Watrin, P., & Fairon, C. (2014, May). FLELex: A graded Lexical Resource for French Foreign Learners. In N. C. (Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA).
- François, T., Volodina, E., Pilán, I., & Tack, A. (2016–May 28). SVALex: A CEFR-Graded Lexical Resource for Swedish Foreign and Second Language Learners. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 213–219. http://www.lrec-conf.org/proceedings/lrec2016/pdf/275_Paper.pdf

References IV

- Gala, N., Tack, A., Javourey-Drevet, L., François, T., & Ziegler, J. C. (2020). Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020).*
- Gooding, S., Kochmar, E., Sarkar, A., & Blackwell, A. (2019). Comparative judgments are more consistent than binary classification for labelling word complexity. *Proceedings of the 13th Linguistic Annotation Workshop*, 208–214. https://doi.org/10/ggdhz3
- Gougenheim, G., Michéa, R., Rivenc, P., & Sauvageot, A. (1964). *L'élaboration du français fondamental (1er degré). Étude sur l'établissement d'un vocabulaire et d'une grammaire de base* [OCLC: 489888938]. Didier.

References V

- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*
- Gunning, R. (1952). *The Technique of Clear Writing*. McGraw-Hill.
- Huckin, T., & Coady, J. (1999). Incidental Vocabulary Acquisition in a Second Language: A Review. *Studies in Second Language Acquisition*, *21*(2), 181–193. https://search.proquest.com/docview/85691136?accountid=17215
- Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 258–286). Cambridge University Press. https://doi.org/10.1017/CBO9781139524780

References VI

- Jeon, E. H., & Yamashita, J. (2014). L2 Reading Comprehension and Its Correlates: A Meta-Analysis [WOS:000331197200006]. *Language Learning*, *64*(1), 160–212. https://doi.org/10.1111/lang.12034
- Kortmann, B., & Szmrecsanyi, B. (Eds.). (2012). *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact* [OCLC: 797182076]. De Gruyter.
- Krashen, S. (1989). We Acquire Vocabulary and Spelling by Reading: Additional Evidence for the Input Hypothesis. *The Modern Language Journal*, *73*(4), 440–464.
- Lee, J., & Yeung, C. Y. (2018). Automatic prediction of vocabulary knowledge for learners of Chinese as a foreign language [INSPEC:17823870]. *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, 4 pp. https://doi.org/10.1109/ICNLSP.2018.8374392

References VII

Maddela, M., & Xu, W. (2018). A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3749–3760. http://aclweb.org/anthology/D18-1410

Melby-Lervag, M., & Lervag, A. (2014). Reading Comprehension and Its Underlying Components in Second-Language Learners: A Meta-Analysis of Studies Comparing First- and Second-Language Learners [WOS:000331845100003]. *Psychological Bulletin*, *140*(2), 409–433. https://doi.org/10.1037/a0033890

Paetzold, G., & Specia, L. (2016). SemEval 2016 Task 11: Complex Word Identification. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 560–569. Retrieved April 4, 2018, from http://www.aclweb.org/anthology/S16-1085

References VIII

- Paetzold, G. H., & Specia, L. (2016–December 17). Understanding the Lexical Simplification Needs of Non-Native Speakers of English. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 717–727.
- Postma, M. C., Miltenburg, E., Segers, R., Schoen, A., & Vossen, P. T. J. M. (2016). Open Dutch WordNet. *Proceedings of the Eigth Global Wordnet Conference.*
- Raptis, H. (1997). Is second language reading vocabulary best learned by reading? [WOS:000077047800005]. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes*, *53*(3), 566–580.

References IX

Sevens, L., Jacobs, G., Vandeghinste, V., Schuurman, I., & Van Eynde, F. (2016). Improving Text-to-Pictograph Translation Through Word Sense Disambiguation. *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 131–135. https://doi.org/10/gf5npr

Shardlow, M. (2013). The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, 69–77. Retrieved April 7, 2018, from http://www.aclweb.org/anthology/W13-2908

References X

- Štajner, S., Yaneva, V., Mitkov, R., & Ponzetto, S. P. (2017). Effects of Lexical Properties on Viewing Time per Word in Autistic and Neurotypical Readers. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 271–281. Retrieved April 14, 2018, from http://www.aclweb.org/anthology/W17-5030
- Tack, A., François, T., Desmet, P., & Fairon, C. (2018). NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 137–146. Retrieved June 1, 2018, from https://www.aclweb.org/anthology/W18-0514
- Tack, A., François, T., Desmet, P., & Fairon, C. (2020, in preparation). The prediction of lexical competence in foreign language reading: A systematic scoping review.

References XI

- Tharp, J. B. (1939). The Measurement of Vocabulary Difficulty. *The Modern Language Journal*, *24*(3), 169–178. https://doi.org/10.1111/j.1540-4781.1939.tb02893.x
- Vahedi, V. S., Ghonsooly, B., & Pishghadam, R. (2016). Vocabulary Glossing: A Meta-Analysis of the Relative Effectiveness of Different Gloss Types on L2 Vocabulary Acquisition. *Teaching English with Technology*, *16*(2), 3–25. https://search.proquest.com/docview/1895971522?accountid=12156
- van den Bosch, A., Busser, B., Canisius, S., & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. Eynde, P. Dirix, I. Schuurman, & V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting* (pp. 99–114). Retrieved January 17, 2017, from http://ilk.uvt.nl/downloads/pub/papers/tadpole-final.pdf

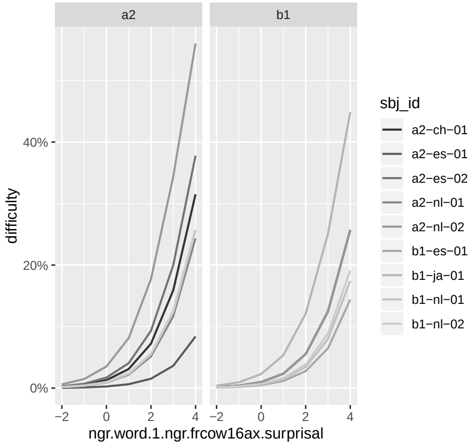
References XII

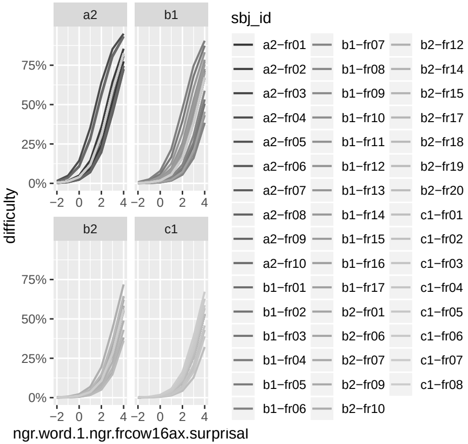
- Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., & Zampieri, M. (2018). A Report on the Complex Word Identification Shared Task 2018. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 66–78. Retrieved June 1, 2018, from https://www.aclweb.org/anthology/W18-0507
- Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis [WOS:000288965700003]. *Computer Assisted Language Learning*, *24*(1), 39–58. https://doi.org/10.1080/09588221.2010.523285

References XIII

Zampieri, M., Malmasi, S., Paetzold, G., & Specia, L. (2017). Complex Word Identification: Challenges in Data Annotation and System Performance. *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, 59–63. Retrieved April 6, 2018, from http://www.aclweb.org/anthology/W17-5910

Random effects of unigram surprisal (Trials 1 vs. 2)



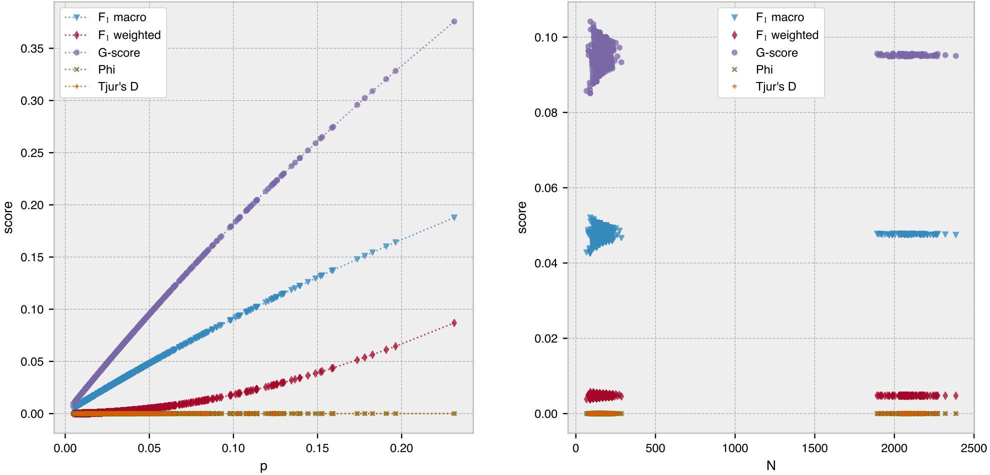


lme4::glmer on FastText word embeddings

	Trial 1				Trial 2			
FASTTEXT	β	*SE*	e^{β}	95% CI	β	*SE*	e^{β}	95% CI
(Intercept)	-6.05 ***	0.25	0.00	[0.00, 0.00]	-5.18 ***	0.27	0.01	[0.00, 0.01]
emb.dim.cc.84	2.87 ***	0.07	17.67	[15.48, 20.15]	1.23 ***	0.07	3.43	[2.99, 3.94]
emb.dim.cc.100	-1.90 ***	0.06	0.15	[0.13 , 0.17]	-1.58 ***	0.08	0.21	[0.17 , 0.24]
emb.dim.cc.16	0.64 ***	0.02	1.91	[1.84, 1.97]	0.81 ***	0.03	2.24	[2.12 , 2.37]
emb.dim.cc.154					0.85 ***	0.04	2.33	[2.14 , 2.53]
τ_{00} pro_level:sbj_id	0.52				0.40			
τ_{00} pro_level					0.24			
ICC	0.14				0.16			
Marginal R^2	0.84				0.69			
Conditional R^2	0.86				0.74			

*** $p < .001$

Sensitivity analysis on 10-fold cross-validation test sets



Performance on CWI benchmarks

Table: Performance on the first CWI shared task (G. Paetzold & Specia, 2016)

System	Type	Accuracy	Precision	Recall	F_1	G-score	ϕ	D
DNN	neural	0.85	0.17	0.58	0.26	0.69	**0.25**	**0.33**
SV000gg	ensemble	0.78	0.15	0.77	0.25	0.77	—	—
CoastalCPH	neural	0.69	0.06	0.40	0.11	0.51	—	—