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
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
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

<table>
<thead>
<tr>
<th>Threshold methods</th>
<th>word length (e.g., $\geq 3$ syllables) (Gunning, 1952)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexicon methods</td>
<td>word frequency (e.g., basic vocabulary list)</td>
</tr>
<tr>
<td></td>
<td>(Gougenheim et al., 1964)</td>
</tr>
<tr>
<td>Empirical &amp; statistical methods</td>
<td>observe from learner data</td>
</tr>
<tr>
<td></td>
<td>engineer features, machine learning</td>
</tr>
</tbody>
</table>
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
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
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
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?
Predicting lexical competence in L2 reading

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
Predicting lexical competence in L2 reading

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
Predicting lexical competence in L2 reading

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
Predicting lexical competence in L2 reading

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
Predicting lexical competence in L2 reading

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

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

- **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)

<table>
<thead>
<tr>
<th>System</th>
<th>Type</th>
<th>F₁</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV000gg</td>
<td>ensemble</td>
<td>0.25</td>
<td>0.77</td>
</tr>
<tr>
<td>PLUJAGH</td>
<td>threshold</td>
<td>0.35</td>
<td>0.61</td>
</tr>
<tr>
<td>CoastalCPH</td>
<td>neural</td>
<td>0.11</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Second CWI Shared Task (Yimam et al., 2018)

<table>
<thead>
<tr>
<th>Both</th>
<th>China and the Philippines</th>
<th>flexed</th>
<th>their muscles</th>
<th>on Wednesday</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple</td>
<td>simple complex</td>
<td>complex</td>
<td>simple</td>
<td>simple</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0 0.4 0.25 0.0 0.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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
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
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
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
A priori knowledge of lexical difficulty

CEFR-graded word frequencies

L2 textbooks readers

A1  A2  B1  B2  C1  C2  OOV

out-of-vocabulary words

CEFRLex project

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

**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)
### NT2Lex, a graded lexicon linked to Open Dutch WordNet

The added value of word-sense disambiguation

<table>
<thead>
<tr>
<th>lemma</th>
<th>part of speech</th>
<th>sense</th>
<th>gloss</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o semantic disambiguation</td>
<td>15,227 entries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>omgangstaal</td>
<td>n</td>
<td>?</td>
<td>'vernacular'</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pakken</td>
<td>v</td>
<td>?</td>
<td>'grab, ...'</td>
<td>708</td>
<td>685</td>
<td>398</td>
<td>19</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>lemma</th>
<th>part of speech</th>
<th>sense</th>
<th>gloss</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/ semantic disambiguation</td>
<td>17,743 entries</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>omgangstaal</td>
<td>n</td>
<td>1</td>
<td>'vernacular'</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pakken</td>
<td>v</td>
<td>1</td>
<td>'grab'</td>
<td>35</td>
<td>117</td>
<td>101</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
Index of semantic complexity

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

**pakken** (verb)

\[ d = .17 \quad \text{get into one’s hands, take physically} \]

\[ d = .83 \quad \text{put at a disadvantage; hinder, harm} \]

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

- overestimates complexity of basic words \((H_{A1} = 7.27, p = .007)\)
- underestimates complexity of advanced words \((H_{C1} = 6.91, p = .009)\)
CEFR-graded word frequencies

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

- **resource = FLELex**
  - Level: A1, A2, B1, B2, C1, C2
  - Type: etymology, translation

- **resource = NT2Lex**
  - Level: A1, A2, B1, B2, C1, C2
  - Type: etymology, translation

A. Tack (CENTAL ITEC FNRS) Predicting the difficulty of words for L2 readers 30/04/2020 29/71
Lexicon-based identification of difficult words

- 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

CEFR-graded word frequencies

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

Noticing difficulty in self-paced reading
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.

<table>
<thead>
<tr>
<th>Trial 1</th>
<th>Trial 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 subjects (A2/B1, CH/ES/JA/NL)</td>
<td>47 subjects (A2/B1/B2/C1)</td>
</tr>
<tr>
<td>51 texts (extensive reading)</td>
<td>5 texts (different per level)</td>
</tr>
</tbody>
</table>
Observing difficulty as a binomial distribution

\[ y \sim B(N, p) \]

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Lang.</th>
<th>Pop.</th>
<th>N</th>
<th>p</th>
<th>Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( M_p) (SD)</td>
</tr>
<tr>
<td>CWI2016 (test)</td>
<td>EN</td>
<td>L2</td>
<td>88221</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EN</td>
<td>L1 + L2</td>
<td>2095</td>
<td>0.383</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EN</td>
<td>L1 + L2</td>
<td>1287</td>
<td>0.424</td>
<td></td>
</tr>
<tr>
<td>CWI2018 (test)</td>
<td>EN</td>
<td>L1 + L2</td>
<td>870</td>
<td>0.505</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DE</td>
<td>L1 + L2</td>
<td>959</td>
<td>0.392</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ES</td>
<td>L1 + L2</td>
<td>2233</td>
<td>0.406</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FR</td>
<td>L1 + L2</td>
<td>2251</td>
<td>0.292</td>
<td></td>
</tr>
<tr>
<td>Trial 1</td>
<td>FR</td>
<td>L2</td>
<td>189084</td>
<td>0.053</td>
<td>0.053 ± 0.036</td>
</tr>
<tr>
<td>Trial 2</td>
<td>FR</td>
<td>L2</td>
<td>72970</td>
<td>0.040</td>
<td>0.041 ± 0.032</td>
</tr>
</tbody>
</table>
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
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
The need to take into account random effects

<table>
<thead>
<tr>
<th>Trial 1</th>
<th>Trial 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL MODEL</td>
<td></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>$\beta$</td>
</tr>
<tr>
<td></td>
<td>$-3.07^{***}$</td>
</tr>
<tr>
<td>$\tau_{00\text{ pro_level:subj_id}}$</td>
<td>0.45</td>
</tr>
<tr>
<td>$\tau_{00\text{ pro_level}}$</td>
<td>0.12</td>
</tr>
<tr>
<td>ICC</td>
<td></td>
</tr>
</tbody>
</table>

*** $p < .001$
Generalized linear mixed-effects analysis

Model selection

\[
\ln \left[ \frac{P(Y_{ij} = 1 \mid u_{0j})}{P(Y_{ij} = 0 \mid u_{0j})} \right] = \beta_0j + \beta_1jX_{1ij} + \ldots + \beta_mjX_{mij}
\]

\[
\beta_0j = \gamma_{00} + \gamma_{01}Z_{1j} + \ldots + \gamma_{0n}Z_{nj} + u_{0j}
\]

\[
\beta_mj = \gamma_{m0}
\]

**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 \((\geq 4)\)  
▶ stepwise forward selection with AIC
### Features of lexical complexity

More than 200 predictors tested

<table>
<thead>
<tr>
<th>Form</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>length of the word (in characters, letters, syllables, stem)</td>
</tr>
<tr>
<td>character</td>
<td>character $n$-grams (sequence likelihood, entropy)</td>
</tr>
<tr>
<td>OLD20 norm</td>
<td>OLD20 norm (average orthographic Levenshtein distance)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Meaning</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>FastText</td>
<td>FastText dimensions, vector similarity</td>
</tr>
<tr>
<td>WordNet</td>
<td>WordNet (hyponymy, hypernymy, synonymy, ...)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Use</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>word $n$-gram</td>
<td>word $n$-gram likelihood in FRCOW16</td>
</tr>
<tr>
<td>morphosyntactic function</td>
<td>morphosyntactic function (part of speech, category)</td>
</tr>
<tr>
<td>syntactic dependency function</td>
<td>syntactic dependency function (depth, head distance, ...)</td>
</tr>
<tr>
<td>frequency</td>
<td>frequency in Manulex and FLELex</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexique3</td>
<td>Lexique3 norms</td>
</tr>
<tr>
<td>order</td>
<td>order of occurrence, exposure, spacing in reading task</td>
</tr>
<tr>
<td>readability</td>
<td>readability thresholds (polysyllables, basic vocabulary)</td>
</tr>
<tr>
<td>etymological information</td>
<td>etymological information (borrowing, cognates, ...)</td>
</tr>
</tbody>
</table>
Generalized linear mixed-effects analysis

### lme4::glmer with random intercepts

<table>
<thead>
<tr>
<th>FULL MODEL</th>
<th>Trial 1</th>
<th>Trial 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>(\beta)  = -4.62***</td>
<td>(\beta)  = -4.08***</td>
</tr>
<tr>
<td></td>
<td>(SE)  = 0.27</td>
<td>(SE)  = 0.32</td>
</tr>
<tr>
<td></td>
<td>(e^\beta)  = 0.01</td>
<td>(e^\beta)  = 0.02</td>
</tr>
<tr>
<td></td>
<td>95% CI [0.01, 0.02]</td>
<td>95% CI [0.01, 0.03]</td>
</tr>
<tr>
<td>ngr.word.lngr.frcow16ax surprisal</td>
<td>(\beta)  = 0.89***</td>
<td>(\beta)  = 1.18***</td>
</tr>
<tr>
<td></td>
<td>(SE)  = 0.02</td>
<td>(SE)  = 0.03</td>
</tr>
<tr>
<td></td>
<td>(e^\beta)  = 2.43</td>
<td>(e^\beta)  = 3.26</td>
</tr>
<tr>
<td></td>
<td>95% CI [2.36, 2.51]</td>
<td>95% CI [3.07, 3.46]</td>
</tr>
<tr>
<td>rea.list.stopwords</td>
<td>(\beta)  = -1.48***</td>
<td>(\beta)  = -0.45***</td>
</tr>
<tr>
<td></td>
<td>(SE)  = 0.06</td>
<td>(SE)  = 0.07</td>
</tr>
<tr>
<td></td>
<td>(e^\beta)  = 0.23</td>
<td>(e^\beta)  = 0.64</td>
</tr>
<tr>
<td></td>
<td>95% CI [0.20, 0.26]</td>
<td>95% CI [0.56, 0.73]</td>
</tr>
<tr>
<td>occ.expo.docu.l</td>
<td>(\beta)  = -3.74***</td>
<td>(\beta)  = -0.77***</td>
</tr>
<tr>
<td></td>
<td>(SE)  = 0.20</td>
<td>(SE)  = 0.04</td>
</tr>
<tr>
<td></td>
<td>(e^\beta)  = 0.02</td>
<td>(e^\beta)  = 0.46</td>
</tr>
<tr>
<td></td>
<td>95% CI [0.02, 0.04]</td>
<td>95% CI [0.42, 0.49]</td>
</tr>
<tr>
<td>res.FLE Lex-TT.A1_SFI</td>
<td>(\beta)  = -0.38***</td>
<td>(\beta)  = -0.77***</td>
</tr>
<tr>
<td></td>
<td>(SE)  = 0.02</td>
<td>(SE)  = 0.04</td>
</tr>
<tr>
<td></td>
<td>(e^\beta)  = 0.69</td>
<td>(e^\beta)  = 0.46</td>
</tr>
<tr>
<td></td>
<td>95% CI [0.66, 0.71]</td>
<td>95% CI [0.42, 0.49]</td>
</tr>
<tr>
<td>msy.categ_v</td>
<td>(\beta)  = 0.35***</td>
<td>(\beta)  = -0.23***</td>
</tr>
<tr>
<td></td>
<td>(SE)  = 0.03</td>
<td>(SE)  = 0.03</td>
</tr>
<tr>
<td></td>
<td>(e^\beta)  = 1.42</td>
<td>(e^\beta)  = 0.80</td>
</tr>
<tr>
<td></td>
<td>95% CI [1.34, 1.50]</td>
<td>95% CI [0.75, 0.85]</td>
</tr>
<tr>
<td>msy.categ_a</td>
<td>(\beta)  = 0.18***</td>
<td>(\beta)  = -1.41***</td>
</tr>
<tr>
<td></td>
<td>(SE)  = 0.03</td>
<td>(SE)  = 0.16</td>
</tr>
<tr>
<td></td>
<td>(e^\beta)  = 1.20</td>
<td>(e^\beta)  = 0.24</td>
</tr>
<tr>
<td></td>
<td>95% CI [1.12, 1.28]</td>
<td>95% CI [0.18, 0.33]</td>
</tr>
<tr>
<td>ety.borr</td>
<td>(\beta)  = -0.18***</td>
<td>(\beta)  = -3.17***</td>
</tr>
<tr>
<td></td>
<td>(SE)  = 0.01</td>
<td>(SE)  = 0.11</td>
</tr>
<tr>
<td></td>
<td>(e^\beta)  = 0.83</td>
<td>(e^\beta)  = 0.04</td>
</tr>
<tr>
<td></td>
<td>95% CI [0.81, 0.86]</td>
<td>95% CI [0.03, 0.05]</td>
</tr>
<tr>
<td>msy.categ_r</td>
<td>(\beta)  = -0.29***</td>
<td>(\beta)  = -1.57***</td>
</tr>
<tr>
<td></td>
<td>(SE)  = 0.07</td>
<td>(SE)  = 0.11</td>
</tr>
<tr>
<td></td>
<td>(e^\beta)  = 0.75</td>
<td>(e^\beta)  = 0.21</td>
</tr>
<tr>
<td></td>
<td>95% CI [0.66, 0.85]</td>
<td>95% CI [0.17, 0.26]</td>
</tr>
<tr>
<td>msy.categ_e</td>
<td>(\beta)  = -0.29***</td>
<td>(\beta)  = -1.57***</td>
</tr>
<tr>
<td></td>
<td>(SE)  = 0.07</td>
<td>(SE)  = 0.11</td>
</tr>
<tr>
<td></td>
<td>(e^\beta)  = 0.75</td>
<td>(e^\beta)  = 0.21</td>
</tr>
<tr>
<td></td>
<td>95% CI [0.66, 0.85]</td>
<td>95% CI [0.17, 0.26]</td>
</tr>
<tr>
<td>msy.categ_g</td>
<td>(\beta)  = -0.29***</td>
<td>(\beta)  = -1.57***</td>
</tr>
<tr>
<td></td>
<td>(SE)  = 0.07</td>
<td>(SE)  = 0.11</td>
</tr>
<tr>
<td></td>
<td>(e^\beta)  = 0.75</td>
<td>(e^\beta)  = 0.21</td>
</tr>
<tr>
<td></td>
<td>95% CI [0.66, 0.85]</td>
<td>95% CI [0.17, 0.26]</td>
</tr>
<tr>
<td>res.Manulex.G1_SFI</td>
<td>(\beta)  = 0.24***</td>
<td>(\beta)  = 0.24***</td>
</tr>
<tr>
<td></td>
<td>(SE)  = 0.03</td>
<td>(SE)  = 0.03</td>
</tr>
<tr>
<td></td>
<td>(e^\beta)  = 1.27</td>
<td>(e^\beta)  = 1.27</td>
</tr>
<tr>
<td></td>
<td>95% CI [1.19, 1.35]</td>
<td>95% CI [1.19, 1.35]</td>
</tr>
</tbody>
</table>

*\(\tau_{00}\) pro_level:sbj_id: 0.61  0.51
*\(\tau_{00}\) pro_level: 0.16  0.36
*ICC: 0.21  0.21

<table>
<thead>
<tr>
<th></th>
<th>Marginal (R^2)</th>
<th>Conditional (R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial 1</td>
<td>0.84</td>
<td>0.51</td>
</tr>
<tr>
<td>Trial 2</td>
<td>0.86</td>
<td>0.61</td>
</tr>
</tbody>
</table>

@A. Taar (CENTAL ITEC FNRS)
Generalized linear mixed-effects analysis

Variance explained by 300-dimensional embeddings

<table>
<thead>
<tr>
<th>Feature</th>
<th>Trial 1 $R^2$</th>
<th>Trial 2 $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ngr.word.1.ngr.fcowl6ax.surprisal</td>
<td>0.87</td>
<td>0.90</td>
</tr>
<tr>
<td>rea.list.stopwords</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>occexpo.docu.l</td>
<td>0.27</td>
<td>0.51</td>
</tr>
<tr>
<td>res.FLELex.TT.AI.SFI</td>
<td>0.72</td>
<td>0.70</td>
</tr>
<tr>
<td>msy.categ_v</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>msy.categ_a</td>
<td>0.75</td>
<td>0.48</td>
</tr>
<tr>
<td>ety.borr</td>
<td>0.44</td>
<td>0.48</td>
</tr>
<tr>
<td>msy.categ_r</td>
<td>0.84</td>
<td>0.94</td>
</tr>
<tr>
<td>msy.categ_e</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>msy.categ_g</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>res.Manulex.G1_SFI</td>
<td>0.71</td>
<td></td>
</tr>
</tbody>
</table>

$^T$ 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
Deep learning of lexical difficulty

Baseline Deep Neural Network

Previous work by De Hertog and Tack, 2018

- top-tier performance on EN & ES at CWI2018 shared task
- limitations
  - recent word embeddings
  - contextualized (sequence) learning
  - 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%)
Deep learning of lexical difficulty

Two enhanced Deep Neural Networks

Contextualized DNN

- character embeddings, convolutions
- pre-trained FastText embeddings

Le village dans son puits de rocher n’était pas encore noyé sous la neige.

The village in its rocky shaft was not yet drowned under the snow.
Deep learning of lexical difficulty

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)
Deep learning of lexical difficulty

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 \( p_{\text{base}} = 0.05 \)
Deep learning of lexical difficulty

Stratified 10-fold cross-validation

Table: Friedman repeated measures (median)

<table>
<thead>
<tr>
<th>Ablation</th>
<th>$\phi$</th>
<th>$D$</th>
<th>$\bar{P}_1$ \textsuperscript{a}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MODELS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNN+P</td>
<td>Full</td>
<td>0.38</td>
<td>0.67</td>
</tr>
<tr>
<td>– L1</td>
<td>0.37</td>
<td>0.65</td>
<td>0.81</td>
</tr>
<tr>
<td>– ID</td>
<td>0.35</td>
<td>0.67</td>
<td>0.86</td>
</tr>
<tr>
<td>– Level</td>
<td>0.35</td>
<td>0.63</td>
<td>0.83</td>
</tr>
<tr>
<td>DNN</td>
<td>Full</td>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td>– CharCNN</td>
<td>0.32</td>
<td>0.65</td>
<td>0.87</td>
</tr>
<tr>
<td>– FastText</td>
<td>0.19</td>
<td>0.45</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>BASELINES</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeHertog2018</td>
<td></td>
<td>0.41</td>
<td>0.58</td>
</tr>
<tr>
<td>CEFR</td>
<td>0.12</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Constant</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

\textsuperscript{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
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
## 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?
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


References XI


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
Random effects of unigram surprisal (Trials 1 vs. 2)
# lme4::glmer on FastText word embeddings

<table>
<thead>
<tr>
<th>FASTTEXT</th>
<th>Trial 1</th>
<th></th>
<th></th>
<th>Trial 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>$\beta$ = -6.05 ***</td>
<td>$SE = 0.25$</td>
<td>$e^{\beta} = 0.00$</td>
<td>$95% CI = [0.00, 0.00]$</td>
<td>$\beta$ = -5.18 ***</td>
<td>$SE = 0.27$</td>
</tr>
<tr>
<td>emb.dim.cc.84</td>
<td>2.87 ***</td>
<td>0.07</td>
<td>17.67</td>
<td>[15.48, 20.15]</td>
<td>1.23 ***</td>
<td>0.07</td>
</tr>
<tr>
<td>emb.dim.cc.100</td>
<td>-1.90 ***</td>
<td>0.06</td>
<td>0.15</td>
<td>[0.13, 0.17]</td>
<td>-1.58 ***</td>
<td>0.08</td>
</tr>
<tr>
<td>emb.dim.cc.16</td>
<td>0.64 ***</td>
<td>0.02</td>
<td>1.91</td>
<td>[1.84, 1.97]</td>
<td>0.81 ***</td>
<td>0.03</td>
</tr>
<tr>
<td>emb.dim.cc.154</td>
<td>0.85 ***</td>
<td>0.04</td>
<td>2.33</td>
<td>[2.14, 2.53]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T_{00}$ pro_level:sbj_id</td>
<td>0.52</td>
<td></td>
<td></td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$T_{00}$ pro_level</td>
<td>0.14</td>
<td></td>
<td></td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC</td>
<td></td>
<td></td>
<td></td>
<td>0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marginal $R^2$</td>
<td>0.84</td>
<td></td>
<td></td>
<td>0.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditional $R^2$</td>
<td>0.86</td>
<td></td>
<td></td>
<td>0.74</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**$*** 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)

<table>
<thead>
<tr>
<th>System</th>
<th>Type</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁</th>
<th>G-score</th>
<th>φ</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN neural</td>
<td></td>
<td>0.85</td>
<td>0.17</td>
<td>0.58</td>
<td>0.26</td>
<td>0.69</td>
<td>0.25</td>
<td>0.33</td>
</tr>
<tr>
<td>SV000gg ensemble</td>
<td></td>
<td>0.78</td>
<td>0.15</td>
<td>0.77</td>
<td>0.25</td>
<td>0.77</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CoastalCPH neural</td>
<td></td>
<td>0.69</td>
<td>0.06</td>
<td>0.40</td>
<td>0.11</td>
<td>0.51</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>