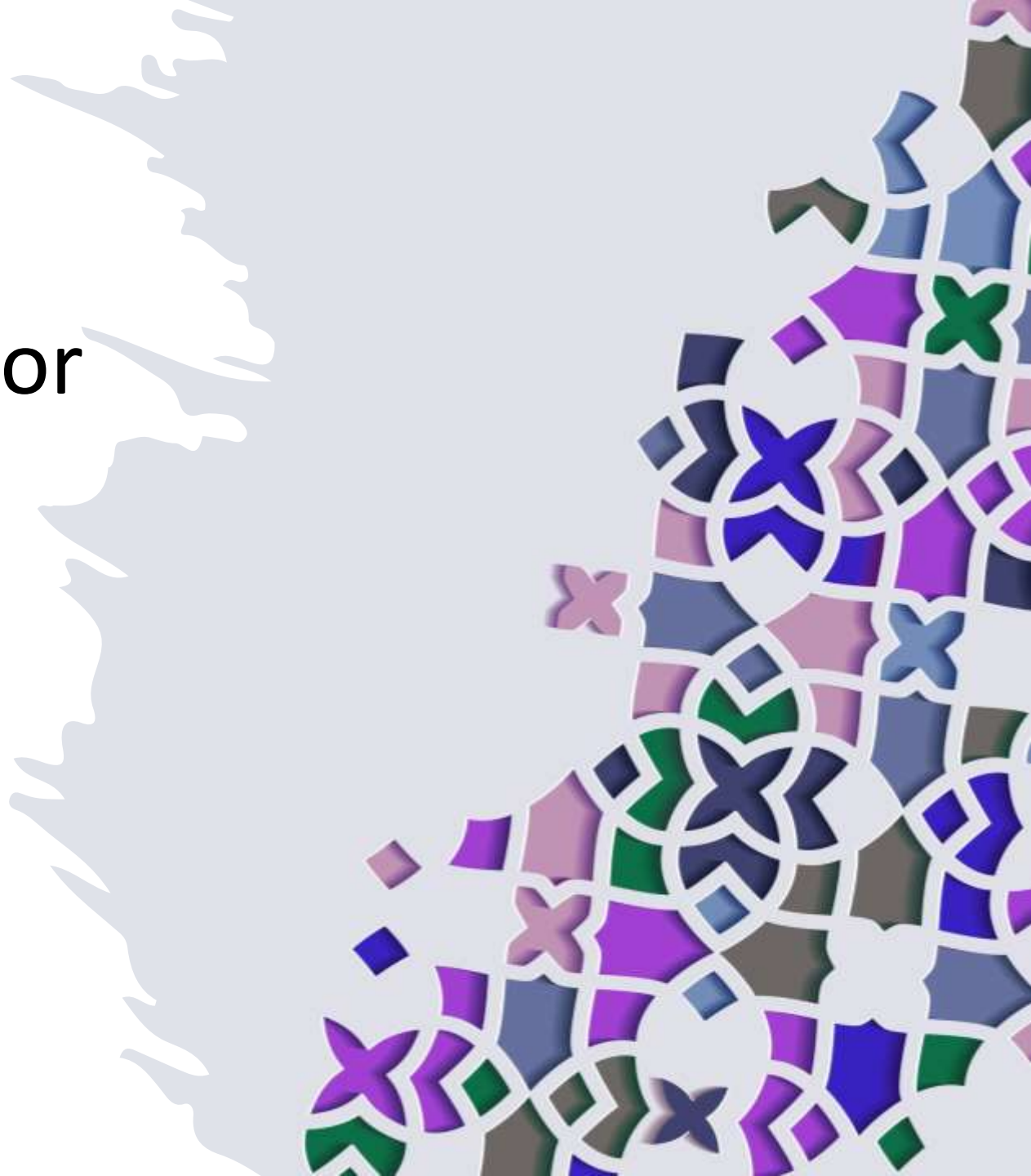
From corpus to profiles: the Icelandic learner error corpus

Isidora Glišić

UNIVERSITY OF ICELAND

Workshop on Profiling second language vocabulary and grammar

Gothenburg, Sweden, 20-21 April 2023



Overview

1. Icelandic as a learner language; Icelandic and CEFR

2. Icelandic error corpora & the first learner corpus

3. Contrastive interlanguage analysis based on CEFR scale

4. Conclusions, challenges, future plans

Icelandic and CEFR

- Demand for teaching Icelandic as SL/FL quite new
  - 2023: around 15–20% of the Icelandic population are immigrants
- No defined CEFR scale for grammar or vocabulary
  - Self-assessment grid from 2001
- No frequency lists for word families
- *Icelandic as a second language* program at the University of Iceland:
  - one-year Practical diploma in Icelandic (est. proficiency level A1-A2)
  - 3-year bachelor degree - students are estimated to be on the level B1-B2 by the end of the first year and reach B2-C1 by the end of the program
- Growing and urgent need for creating learner profiles, new teaching materials, standardized testing

Icelandic error corpora

- CLARIN-IS digital resources repository (Database of Modern Icelandic Inflections, Database of Icelandic Morphology, Icelandic Gigaword corpus etc.)
- Government project *Language Technology for Icelandic 2019-2023*
  - First error corpora for Icelandic:

    The Icelandic Error Corpus (IceEC)

    Three specialized error corpora (2019-2022): The Icelandic L2 Error Corpus, The Icelandic Dyslexia Error Corpus, and The Icelandic Child Language Error Corpus

Corpora	Number of words	Errors per 1000/w
General	1,239,024	45.67
L2	162,071	153.93
Children	37,443	208.77
Dyslexia	38,891	216.91

Creating the corpora and annotation scheme

```
5990  <w>gera</w>
5991  <w>ráð</w>
5992  <w>fyrir</w>
5993  <revision id="255">
5994    <original><w>tvö</w><w>myndbrigði</w><w>fyrir</w></original>
5995    <corrected><w>tveim</w><w>myndbrigðum</w><w>í</w></corrected>
5996      <errors>
5997        <error xtype="case-collocation" idx="255-1" eid="0" />
5998        <error xtype="nominal-inflection" idx="255-1" eid="0" />
5999        <error xtype="wrong-prep" idx="255-2" eid="0" />
6000      </errors>
6001  </revision>
6002  <w>þágufallsendingu</w>
```

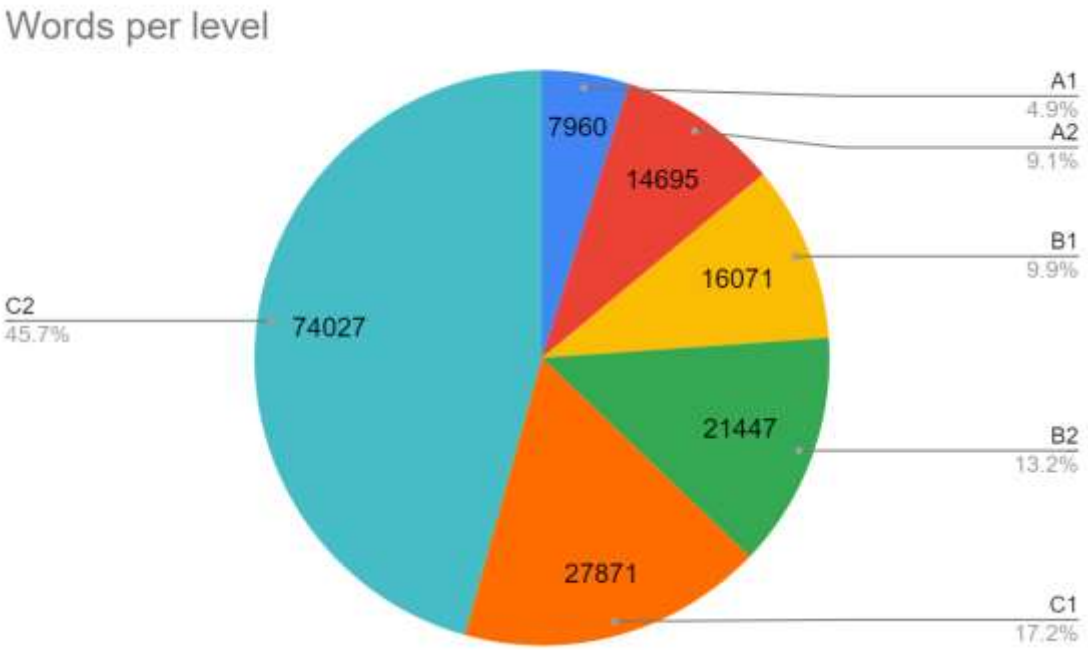
An example of revision spans with multiple error codes.

- The steps involved:
  - gathering large quantities of texts within each category, manually proofreading and marking errors (each error manually labeled within a pre-formed annotation scheme)
  - published in augmented TEI-format XML documents. Specific TEI element, *revision*, was created to map out the differences between the original text file and the corrected file.

- Error classification scheme combined linguistic errors (morphology, syntax, etc.) and surface structure taxonomies (omission, addition, etc.)
- 5 main categories (orthography, grammar, vocabulary, coherence, and style), each further divided into more descriptive subcategories, divided into error codes - 258 in total
  - general: *wording*
  - specific: *af4að*, *i4y*

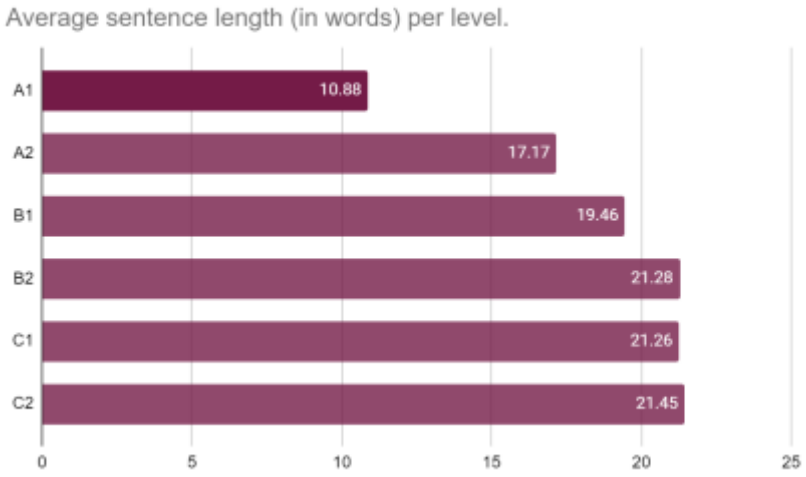
The Icelandic L2 Error Corpus (IceL2EC)

- Text collection: public online submission form (texts previously unpublished and obtained directly from their authors)
- Original annotation scheme expanded with new labels that were specific to the L2 errors
- Metadata: author's first language, other languages, length of residence in Iceland, length of study of Icelandic, and ***proficiency level***



Number of revision: 17241
Number of errors: 24948
Number of files: 101
Number of words: 162071
Errors p/1000 words: 153.93
Average words per file: 1605

17 different L1s

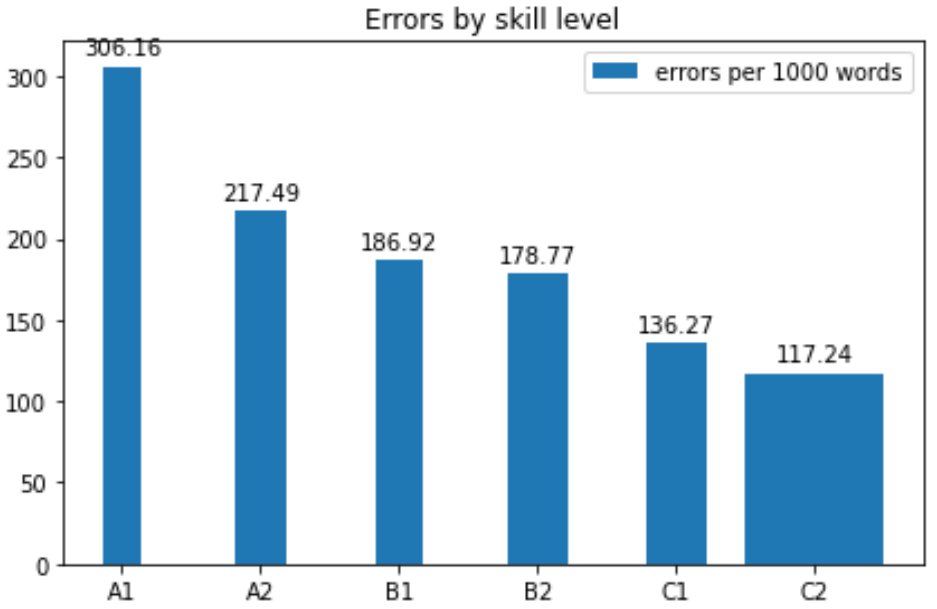


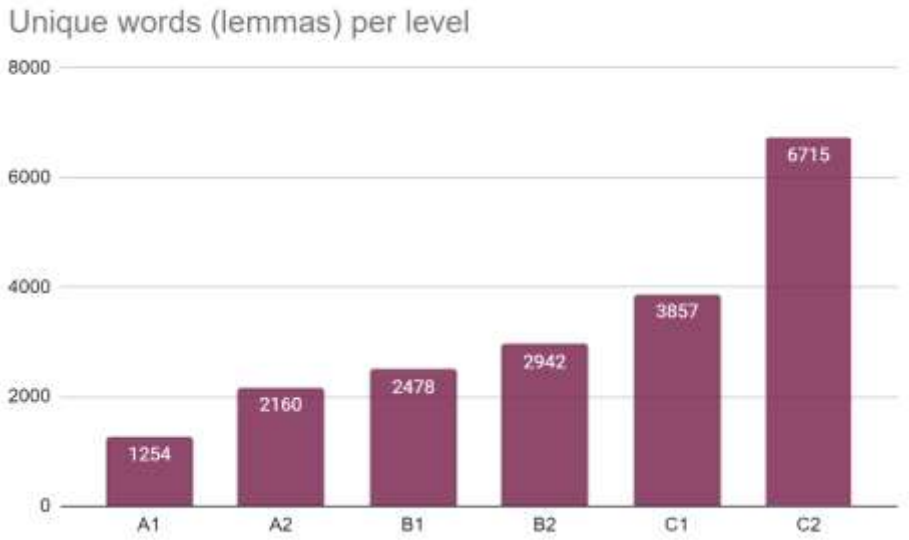
Word and error frequency

- Most frequent errors within grammar (43.57%) – mostly involve case government, agreement, and definiteness
  (grammar category accounts for only 11.8% in the general Icelandic Error Corpus)
- Example: 5 most common errors for level A2 (%)

wording	19.94
nominal-inflection	14.22
wrong-prep	7.83
context	7.42
case-prep	6.87

- Unique lemmas total: 12167
  663 that appear on A1 but not on A2
  1569 that appear on A2 but not on A1
- Corrected versions of texts used





Future steps

- Icelandic Error Corpus is already used in creating a grammar and spelling correction software - *GreynirCorrect* (only open-source Icelandic spell checker and a part of the spelling and correction module of the Language Technology Program)
- Using the L2 corpus in a collaborative project on mapping out the grammar and vocabulary on the CEFR scale
- Train a machine learning model for automatic skill level detection based on extracted features

- Issues: scarcity and inconsistency of data, PoS taggers, manual error tagging vs. automatic error detection

References

Arnardóttir, Þ., Xu, X., Guðmundsdóttir, D., Stefánsdóttir, L. B. and Ingason, A. K. (2021). Creating an Error Corpus: Annotation and Applicability. In Proceedings of CLARIN 2021, Monica Monachini and Maria Eskevich (eds.), pp. 59-63.

Arnardóttir, Þ., Glisic, I., Simonsen, A., and Stefánsdóttir, L. (2022). Error Corpora for Different Informant Groups:Annotating and Analyzing Texts from L2 Speakers, People with Dyslexia and Children. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 245–252, New Delhi, India.

Glisic, I., Ingason, A.K. (2022). The Nature of Icelandic as a Second Language: An Insight from the Learner Error Corpus for Icelandic. In Proceedings of Linköping Electronic Conference, pp. 23-33.

Garðarsdóttir, M. & Þorvaldsdóttir, S. (2020). A Processability approach to the development of case in L2 Icelandic. Special Issue on L2 Case & Agreement, Language, Interaction and Acquisition.

Kerz, E. & Wiechmann, D. & Qiao, Y. & Tseng, E. & Ströbel, M. (2021). Automated Classification of Written Proficiency Levels on the CEFR-Scale through Complexity Contours and RNNs. In Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 199–209.

MacDonald, P., García-Carbonell, A., Carot-Sierra, J. (2013). Computer learner corpora: Analysing interlanguage errors in synchronous and asynchronous communication. In Language Learning Technology, 17(2), pp. 36–56. Retrieved from http://llt.msu.edu/issues/june2013/macdonaldetal.pdf

Óladóttir, H., Arnardóttir, Þ., Ingason, A. K., and Þorsteinsson, V. (2022). Developing a Spell and Grammar Checker for Icelandic using an Error Corpus. In Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France, pp. 4644-4653.

Thewissen, J. (2013). Capturing l2 accuracy developmental patterns: Insights from an errortagged efl learner corpus. The Modern Language Journal, 97(S1), pp.77–101. doi: https://doi.org/https://doi.org/10.1111/j.1540-4781.2012.01422.x

- **Corpora:**

Ingason, A. K., Arnardóttir, Þ., Stefánsdóttir, L. B., and Xu, X. (2021). The Icelandic Child Language Error Corpus (IceCLEC) version 1.1. CLARIN-IS. https://repository.clarin.is/repository/xmlui/handle/20.500.12537/133

Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., and Xu, X. (2021). Icelandic Error Corpus (IceEC) version 1.1. CLARIN-IS. https://repository.clarin.is/repository/xmlui/handle/20.500.12537/105

Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., Xu, X., Glišic, I., and Guðmundsdóttir, D. (2022). The Icelandic L2 Error Corpus (IceL2EC) version 1.3. CLARIN-IS. https://repository.clarin.is/repository/xmlui/handle/20.500.12537/280

Ingason, A. K., Arnardóttir, Þ., Stefánsdóttir, L. B., Xu, X., Guðmundsdóttir, D., and Glišic, I. (2022). The Icelandic Dyslexia Error Corpus (IceDEC) version 1.2. CLARIN-IS. https://repository.clarin.is/repository/xmlui/handle/20.500.12537/281