The nature of Icelandic as a second language: An insight from the learner error corpus for Icelandic

Isidora Glišić
Anton Karl Ingason

University of Iceland

CLARIN 2021

27 – 29 September 2021

Introduction

Structure

1. Background on examining Icelandic as learner language
   - The Icelandic L2 Error Corpus (IceL2EC)
2. Methods used for: text collection
   text analysis and error classification
3. Data analysis
   - Contrastive interlanguage analysis (CIA) (Granger, 2017)
     - comparing learner language with native speaker reference corpora (L2 vs. L1)
     - comparing different varieties of learner language (L2 vs. L2)
4. Preliminary results

Background

- The novelty of Icelandic as L2
  - first contrastive analysis of the learner language emerged in the 1980s (Sigmundsson, 1987)
- Recent focus on learner language at the University of Iceland (Þorvaldsdóttir and Garðarsdóttir, 2013; Ólafsson, 2016)
- Error corpus for Icelandic as a second language (Ingason et al., 2021)
  - Part of the project *Language Technology for Icelandic 2019-2023 / Specialized error corpora*
  - Still in development

Current state
70 student essays annotated for errors
27 adult second language speakers of Icelandic with 13 different first languages
12081 revision spans and 17508 error instances
Words per text mean: 1780 Errors per 1000 words: 140.73

Methods

- Text collection
  - Public online submission form – texts previously unpublished and obtained directly from their authors
  - Manually proofread and annotated for errors
  - Converted to augmented TEI format XML document with labeled enumerated sentences, words and punctuation, and revision spans with unique id numbers containing errors
  - Annotation system for error labeling originally developed for the Icelandic Error Corpus (Ingason et al., 2020) - later expanded with new labels that were specific to the L2 errors
  - Error tagset: 19 error categories further divided into subclasses, 259 error codes in total
  - Metadata: author's first language, other languages, length of residence in Iceland, length of study of Icelandic, and proficiency level

Methods


```
<w>sínum</w>
<revision id="15">
  <original><c>,</c><w>hópurinn</w><w>springur</w><c>,</c><w>og</w><w>svo</w><w>leitt</w></original>
  <corrected><w>sundrast</w><w>hópurinn</w><c>,</c><w>og</w></corrected>
    <errors>
      <error xtype="extra-comma" idx="15-1" eid="0" />
      <error xtype="wording" idx="15-2" eid="0" />
      <error xtype="v3" idx="15-2" eid="0" />
      <error xtype="wording" idx="15-3" eid="0" />
    </errors>
</revision>
<w>annar</w>
<revision id="16">
  <original><w>hlutinn</w><w>af</w><w>sögunni</w></original>
  <corrected><w>hluti</w><w>sögunnar</w><w>leiðir</w></corrected>
    <errors>
      <error xtype="def4ind" idx="16-1" eid="0" />
      <error xtype="wording" idx="16-2" eid="0" />
      <error xtype="dep" depId="15-3" eid="0" />
    </errors>
</revision>
<w>til</w>
<w>harmleiks</w>
<c>.</c>
```

Example of revision spans with multiple error codes and dependent error

Data analysis

L2 vs. L1

Corpus	Files	Total words	Revisions	Categorized Errors	Errors/1000w
Icelandic Error Corpus	4046	1137941	44261	55346	44.56
Icelandic L2 Error Corpus	70	124626	12081	17508	140.73

- disparity in frequency of error categories and subcategories in L2 Icelandic compared to L1 errors
- most frequent in the L2 corpus: grammar (43.57%), punctuation (12.14%) and wording (11.63%) (grammar category accounts for only 11.8% in the general Icelandic Error Corpus
- 35 error codes that appear only in the L2 corpus, 27 within the grammar category

Data analysis

L2 vs. L1

- Error codes with most similar and most different rankings between the corpora:

SIMILAR			
Error code	**Rank L1**	**Rank L2**	**Δ rank**
wording (wording)	1	1	0
nonword (nonword)	3	3	0
date-period (punctuation)	99	99	0
extra-conjunction (punctuation)	32	33	1
comma4colon (punctuation)	89	90	1

DIFFERENT			
Error code	**Rank L1**	**Rank L2**	**Δ rank**
context (lexical)	121	6	115
case-prep (grammar)	121	9	112
case-verb (grammar)	121	15	106
missing-prep (omission)	120	17	103
extra-prep (insertion)	121	18	103

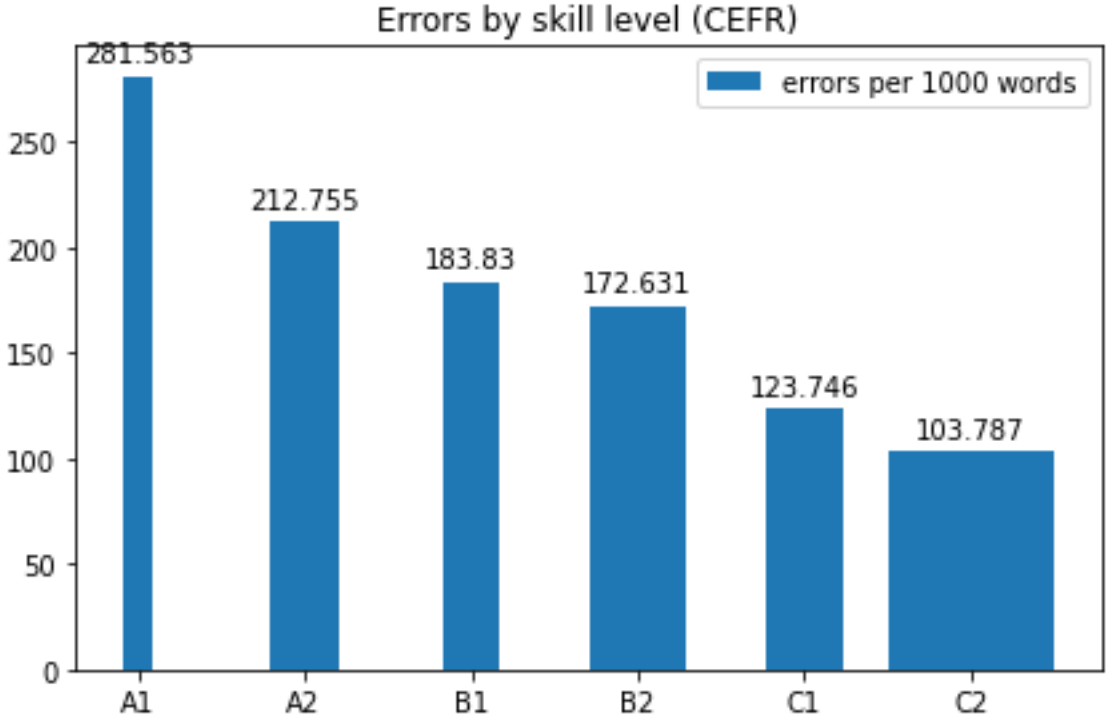
Data analysis

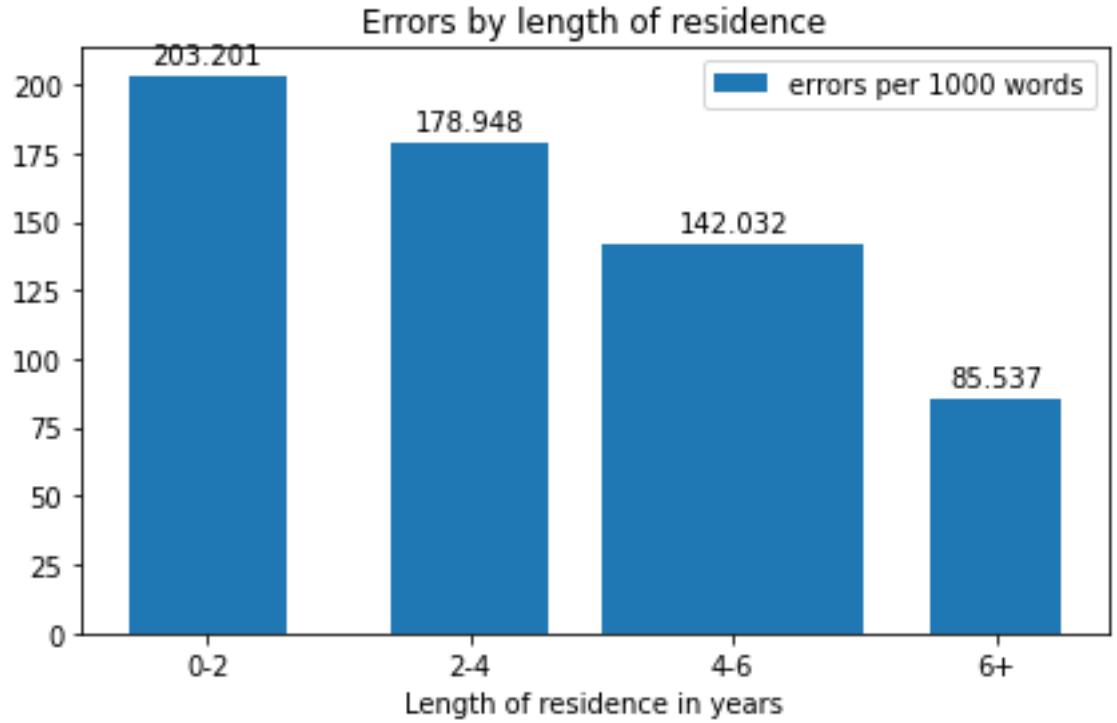
L2 vs. L2

- focus so far on skill level and length of residence
- Total number of files, words, errors and errors per 1000 words per skill level:

Level	Files	Total words	Total errors	Errors/1000w
A1	12	3889	1095	281.56
A2	19	11901	2532	212.76
B1	8	10439	1919	183.83
B2	11	19504	3367	172.63
C1	9	21940	2715	123.75
C2	11	56953	2911	103.79

Data analysis





Conclusions

- Corpus still in development
- Further data analysis underway, focus on L2 vs. L2 CIA
- Corpus still small – individual features difficult to highlight due to the limited sample size
  - Expanding the corpus will provide further possibilities to analyse various demographic and other features of learner language such as first language influence
  - Possible applications: perfecting teaching materials (both electronic, textbooks and syllabi) and automatic correction tools for Icelandic

Resources

Garðarsdóttir, M. and Þorvaldsdóttir, S. 2020. A processability approach to the development of case in L2 Icelandic. *Language, Interaction and Acquisition A cross-theoretical and cross-linguistic perspective on the L2 acquisition of case systems*, 11(1):68–98.

Granger, S., 2017.*Learner Corpora in Foreign Language Education*, pages 427–440. Springer International Publishing, Cham.

Ingason, A. K., Stefánsdóttir, L. B., and Arnardóttir, Þ. 2020. Icelandic error corpus (IceEC) version 0.9. CLARIN-IS.

Ingason, A. K., Stefánsdóttir, L. B., Arnardóttir, Þ., Xu, X., and Glišic, I. 2021. The Icelandic L2 error corpus (IceL2EC) version 1.1. CLARIN-IS.

Piccardo, E., Goodier, T., and North, B. 2018. Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Strasbourg: Council of Europe Publishing., 01.

Sigmundsson, S. 1987. Íslenska í samanburði við önnur mál. *Íslenskt mál og almenn málfræði*, 9.

Thewissen, J. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(S1):77–101.

Ólafsson, G. H. 2016. Grammar and linguistic structures at level A1 of Icelandic. Master's thesis, University of Iceland, Unpublished, 6.

Þorvaldsdóttir, S. and Garðarsdóttir, M. 2013. Fallatileinkun í íslensku sem öðru máli. *Milli mála: Tímarit um erlend tungumál og menningu*, 5:45–73.