Context-sensitive proofreading for a minority language

Anton Karl Ingason

Department of Icelandic, University of Iceland, Reykjavík

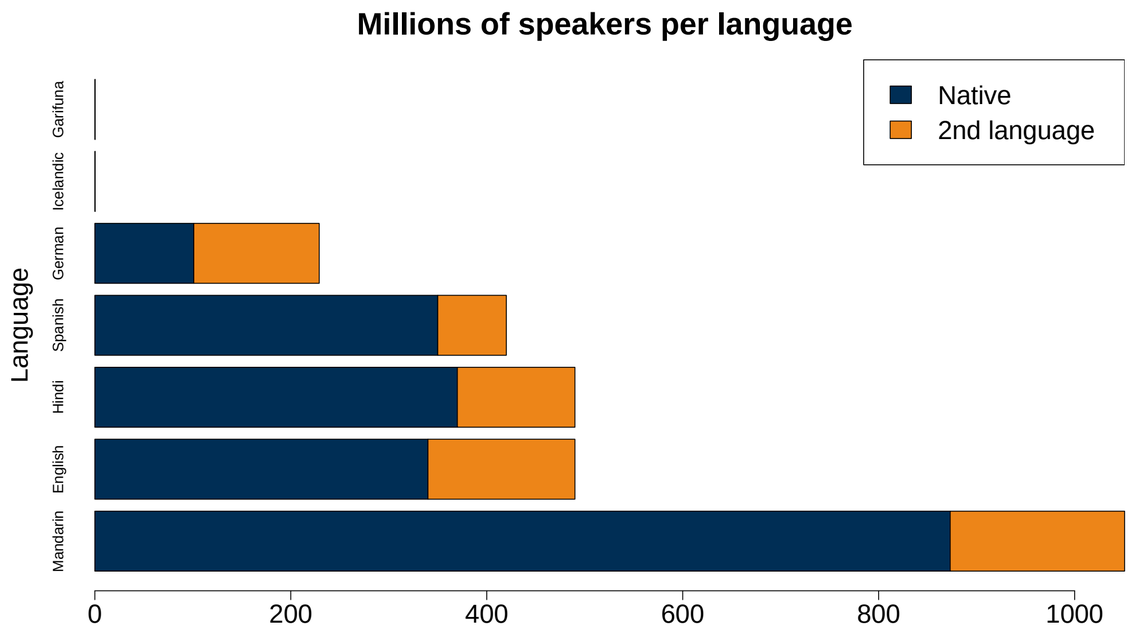
UNIVERSITY OF ICELAND

Introduction

- From the point of view of language checking software
  (and Language Technology (LT) in general)
  ... most languages are, in fact, minority languages
  ... since established solutions may not apply without modification
- Problems associated with small languages and proofreading
  - Lack of resources (people, money)
  - Typological difference from the larger languages (morphology)
- The current proposal for proofreading software:
  - Make use of language-independent solutions
  - Combined with the advantages of Free and Open Source projects
- I present a prototype based on LanguageTool (LTool)
  - This is work in progress!
  - Thanks to LTool integration – OpenOffice.org is supported
  - **Goal:** Practical and viable language checking for Icelandic

Icelandic as a Minority Language in the LT Context

- Global aspects of proofreading software target the languages of 6,803,000,000 people
- Language specific aspects target very different markets
- Icelandic ≈ 320.000 speakers
- Garifuna ≈ 300.000 speakers (Ravindranath 2009)



The Icelandic Situation

- Weaknesses
  - Tiny market and limited resources
  - Morphological richness beyond what we get for the largest languages (standard POS-tagset has about 700 different tags)
- Strengths
  - The IceNLP toolkit exists, with a POS-tagger, a shallow parser, a lemmatizer and some more tools. All LGPL-licensed. (http://sourceforge.net/projects/icenlp/)

Our Previous Approach (Ingason et al. 2009)

- Experiments were carried out using a language independent data-driven method in the spirit of Golding (1995)
  - Machine-Learning-based disambiguation of confusion sets, e.g. $C = \{\textit{pear}, \textit{pair}\}$ in "a nice pear of shoes"
  - Features extracted from word context included information about the words in the context: word forms, lemmas and plenty of morphosyntactic features (case, gender, number, mood, etc.)
- Despite a high number of features extracted to cope with the morphological richness of the language – results were worse than in experiments for English, 80.9%–87.2% precision, depending on the classification algorithm, compared to over 90% for English
- Not so practical – doesn't scale nicely

```
48    <example type="correct">Breytingar eru á næsta <marker>leiti</marker>.</example>
49  </rule>
50  <rulegroup id="LEYTI" name="leyti">
51    <rule>
52      <pattern mark_from="2" mark_to="0">
53        <token>að</token>
54        <token regexp="yes">sumu|ýmsu|verulegu|einhverju|ákveðnu|öllu|þessu|því|flestu|miklu</token>
55        <token>leiti</token>
56      </pattern>
57      <message>Skrifa skal 'y' í orðasambandinu: að <match no="2"/> <suggestion>leyti</suggestion></message>
58      <short>Skrifa 'y' í: að [...] leyti</short>
59      <example type="incorrect">Petta er rétt að sumu <marker>leiti</marker>.</example>
60      <example type="correct">Petta er rétt að sumu <marker>leyti</marker>.</example>
61    </rule>
62    <rule>
63      <pattern mark_from="2" mark_to="0">
64        <token>um</token>
65        <token regexp="yes">það|þetta|svipað|álíka</token>
66        <token>leiti</token>
67      </pattern>
68      <message>Skrifa skal 'y' þegar leyti vísar til tíma eins og í orðasambandinu: um <match no="2"/>
        <suggestion>leyti</suggestion>.</message>
69      <short>Skrifa skal 'y' í: um [...] leyti</short>
70      <example type="incorrect">Við mætum um svipað <marker>leiti</marker> og þið.</example>
```

LanguageTool (www.languagetool.org)

- LTool is a rule-based open source framework for developing various kinds of context-sensitive language checking, including spellchecking (Naber 2003, Milkowski 2010)
- LTool allows us to focus on writing language specific rules for Icelandic using a simple but powerful XML-syntax
- Integration with user software is developed by others
- 20 languages are already supported and developed by the LTool community – including our Icelandic prototype
- Belarusian, Catalan, Danish, Dutch, English, French, Galician, German, **Icelandic**, Italian, Lithuanian, Malayalam, Polish, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, Ukrainian
- **... no Garifuna ... yet!**

Current Status

- Prototype includes 40 correction patterns
- Combining LTool pattern matching with Regular Expressions makes each rule cover a variety of context-sensitive cases
  - Spellchecking
  - Grammar checking
  - Stylistic suggestions
- The focus of the project has evolved from a previous emphasis on Machine Learning to a more practical approach that can be easily extended by a community of users who do not need to have expertise in computer science

Conclusion

- Development of context-sensitive language checking does not need to involve advanced technical expertise or extensive resources
- Thanks to the all-open-source approach and LTool we have a viable project that can be easily extended by others
- For a small language – being able to develop an LT solution with limited resources can be a deciding factor of whether such development occurs at all for the language
- One person can add a new language to LTool (and they can get assistance from other project members)
- The LTool community makes sure a variety of front ends are and will be supported, including: OpenOffice.org, Firefox and Thunderbird

References and Further Information

- See **www.linguist.is/papers** for more material, including information on the references cited above
- Visit **www.languagetool.org** to download LanguageTool for your language
- This work was supported in part by the Icelandic Research Fund