IceTaboo: A database of contextually inappropriate words for Icelandic

Agnes Sólmundsdóttir, Lilja Björk Stefánsdóttir, Anton Karl Ingason
University of Iceland

CLARIN 2021, 27-29 September

What is this project about?

- Creating an annotated database of words that are inappropriate or offensive to at least some speakers in some contexts.
  - Classification of offensive words
  - Reasons for offensiveness
    - Sometimes in a nuanced relationship with context
    - Sometimes (even strongly) against our own intuitions about offensiveness
- Goal to provide a basis for further work on the topic in Icelandic
- Current status:
  - 2725 entries (words), published on CLARIN under CC BY 4.0
  - A subset has been integrated into a style module of a functional spelling and grammar checker in a collaboration with an industry partner, software company Miðeind (being used by real users)

Classes of offensive words

- generally inappropriate words
- swear words
- words associated with alcoholism or drug addiction
- disability words
- health related words
- words regarding stupidity
- gendered words (generally, or ones that discriminate against people based on gender)
- nasty adjectives
- offensive profession names
- collocations
- LGBTQIA+ words used inappropriately
- verbs of inappropriate actions
- offensive words related to religion
- offensive descriptions of people's appearance
- words for genitals
- offensive prefixes
- offensive words related to sex
- offensive nationality words (often linked to oppression of some sort)

Words with nuanced relationship with offensiveness

- Inappropriate for children (while not so for adults)
- political terms (may trigger a negative reaction, depending on a person's political views)
- non-offensive (words that are not really offensive but have a nuanced meaning that may make sense to exclude in contexts that strive to remain neutral),
- words with an alternative, non-offensive meaning (included to establish that the offensive counterpart reading is only attested in certain contexts).

Example entry

- **word:** fóstra(roughly: ‘a daycare babysitter’)
- **part-of-speech:** noun
- **code (see classes above):** m (for profession)
- **code2:** (additional classification) NA
- **meaning:** preschool teacher
- **reason for offensiveness:** Now considered an obsolete and degrading term for the profession of preschool teachers, suggesting they are not a profession of educators.
- **additional information (if needed):** NA
- **alternative non-offensive meaning:** NA

Conclusion

- We have created a database on contextually inappropriate words for Icelandic
- Already published to CLARIN and available under CC BY 4.0
- Already used in language correction software developed in partnership with a software company
- Small project that can serve as a basis for further development
- Future work
    - Add more words and classes
    - Refine the annotation
    - Gather feedback using crowdsourcing
    - Apply methods from other projects (for other languages) to enrich our database