# CLARIN Annual Conference 2021

# PROCEEDINGS

Edited by

Monica Monachini, Maria Eskevich

27 – 29 September 2021
Virtual Edition

Please cite as:
Proceedings of CLARIN Annual Conference 2021. Eds. M. Monachini and M. Eskevich.
Virtual Edition, 2021.

# IceTaboo: A database of contextually inappropriate words for Icelandic

**Agnes Sólmundsdóttir**
University of Iceland
ags46@hi.is

**Lilja Björk Stefánsdóttir**
University of Iceland
lbs@hi.is

**Anton Karl Ingason**
University of Iceland
antoni@hi.is

**Abstract**

We present IceTaboo, a database of 2725 words that are inappropriate or offensive to at least some speakers in some contexts. Every word is coded for part of speech, a classification of reasons that trigger a negative reaction among some speakers as well as information about the meaning expressed by the word. The database is released under an open CC BY 4.0 license on CLARIN and it is already being used in the development of an automatic proofreading tool, developed in collaboration with an industry partner in commercial software development. The proofreading tool, itself, is under development in an open repository on Github under an MIT license.

## 1 Introduction

The detection of offensive or contextually inappropriate language can be important for monitoring freely accessible discussion spaces on social media or for helping a user of a word processing system to avoid inappropriate expressions. A variety of resources of methods have been developed to address this challenge (Davidson et al., 2017; Risch et al., 2019; Pitsilis et al., 2018; Wu et al., 2019; Sigurbergsson and Derczynski, 2019; Pitenis et al., 2020; Mubarak et al., 2020; Çöltekin, 2020). One basic kind of a resource that can be integrated into such systems is a database of words that offend readers, in general or in some context. While such a database is not sufficient for detecting all instances of offensive language, it can serve as a useful first step. Furthermore, the development of such a database facilitates the emergence of a classification of the reasons that triggers reactions of this type in a given language.

In the project we present IceTaboo,[1] the Icelandic Taboo database, that was manually compiled at the Language and Technology Lab of the University of Iceland during 2020. IceTaboo contains 2725 words that are inappropriate or offensive to at least some speakers in some contexts. Every word is coded for part of speech, a classification of reasons that trigger a negative reaction among some speakers as well as information about the meaning expressed by the word. The database is released under an open CC BY 4.0 license on CLARIN and it is already being used in the development of an automatic proofreading tool, developed in collaboration with an industry partner in commercial software development. The proofreading tool, itself, is under development in an open repository on Github under an MIT license. Open access availability on CLARIN and an industry-friendly licensing policy ensures that the resource is ready to be used by any software developer that shows interest, thus supporting to the general policies in the current Language Technology Programme for Icelandic (Nikulásdóttir et al., 2020) that aim to make the delivered output as accessible to future development as possible.

## 2 Offensive language and automatic proofreading

Some previous work on offensive language focuses on training classifiers on annotated training data where the units being annotated are utterances labelled as offensive or not and the classification algorithm



[1] Accessible at: https://repository.clarin.is/repository/xmlui/handle/20.500.12537/64

assigns such utterances to an offensiveness category based on some features extracted from the item in question, for example using bag-of-word classification approaches (Kwok and Wang, 2013; Burnap and Williams, 2015).

Challenges arise when basing the classification on specific word forms, because while a racist word like *n****** triggers a reaction of an offensive experience in generally all contexts, a word like *gay* can be used either in a positive context or a negative one, depending on the intentions of the speaker who produces the utterance and the context in which the relevant discussion takes place. Furthermore, the reason for why a reader may feel offended are of various kinds. One sentence may be perceived as hate speech, invoking strong emotions, whereas another one may be perceived as promoting a negative stereotype, also a negative but less intense emotion. Any of these types of incidents are potential circumstances for intervention, such as filtering or tagging potentially offensive posts on social media or by suggesting a more appropriate choice of words to the user of some word processing software.

Detection of offensive language is an under-explored area of Natural Language Processing for Icelandic. Still, in the context of automatic proofreading, previous work has addressed the disambiguation of confusion sets of the *there/their*-type (Ingason et al., 2009; Friðriksdóttir and Ingason, 2020) and methods that predict the appropriate candidate from a confusion set are in principle applicable to the detection of an appropriate vs. inappropriate use of a word that is offensive in only some contexts. Furthermore, general advances in NLP for the language have paved the way for more advanced context analysis, such as the development of deep phrase structure parsing (Jökulsdóttir et al., 2019; Þorsteinsson et al., 2019; Arnardóttir and Ingason, 2020) – including systems released via CLARIN – which has allowed for complex pattern matching that involves syntactic context. Lemmatization system have also made it possible to normalize Icelandic words to their dictionary forms, a non-trivial task in a language with rich morphology (Ingason et al., 2008; Ingólfsdóttir et al., 2019), and a crucial step in methods that derive features from specific words, and thus important for detection of offensive language.

## 3 The IceTaboo resource

The IceTaboo database consists of a list of words in Icelandic that may in some way be considered inappropriate, taboo and/or loaded in use or meaning. These can be words such as; words that are biased against certain minorities (i.e. people of different races, abilities, genders or sexualities), words that are derogatory towards people, unnecessarily gendered, obsolete and so on. The list also includes words that are not very inappropriate but can be considered an unfortunate topic for children or politically loaded in any way. The words are grouped together in categories depending on either their meaning, form or use. Each word has then been marked with a short explanation (in Icelandic) on how they can be considered inappropriate and in what context. The words were collected through brainstorming sessions and systematic follow-ups to these (see below), but other similar lists from other sources were also used, i.e. a list of taboo words for children from the project Samrómur (Mollberg et al., 2020) and a list of taboo words for childrens Scrabble developed by the sofware company Miðeind.

This list does not contain actual information or data on the real opinion of the public towards these words. These words are merely thought to elicit a negative reaction for at least some speakers in at least some context. This list can therefore be a useful data set that serves as a starting point for further analysis. The list is already being used in the development of a commercial automatic proofreading system in collaboration with the software company Miðeind.[2] The database identifies the following classes of offensive words:

(1) **Classes of offensive words:**
Generally inappropriate, 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 either women or men), 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 gen-

[2] See: `https://github.com/mideind/GreynirCorrect`

tials, offensive prefixes, offensive words related to sex, offensive nationality words (often linked to oppression of some sort).

Additional classes:

(2) **Words with nuanced relationships 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 neural), words with an alternative, non-offensive meaning (included to establish that the offensive counterpart reading is only attested in certain contexts).

An example entry is shown below. The word *fóstra* is often considered obsolete and degrading although some members of the profession still prefer to use this word and do not experience it as a negative expression.

- 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

The database was compiled manually, using the creativity of a number of research assistants, followed by a systematic search carried out mostly by the first author, supervised by the other two authors. Brainstorming sessions were held, in which RA's thought of all the most offensive words they could think of and the output of this work was used to establish the classification system. Then, each class of words was systematically studied, looking for synonyms or related words that might belong in the list, and electronic resources were used to look for compounds that contained inappropriate parts.

## 4 Conclusion

In the present paper, we have presented IceTaboo, a novel resource for processing offensive words in Icelandic. Although the work presented here involves manual annotation, and is already being used to highlight inappropriate words in a commercial automatic proofreading system via a simple lookup (after lemmatization), we believe it may also be of use in future development of systems that apply machine learning methods to automatic detection of offensive language. The words in the database can inform feature extraction steps of such systems and potentially make them more effective. We acknowledge that this release is only a small step to that end, yet we believe it is significant and has potential to facilitate further work on the topic in the context of Icelandic.

Various further avenues for future work remain to be explored. As suggested by reviewers, it would be interesting to test these materials on a wider audience to study the reactions of, for instance, younger vs. older speakers. It would also be worth trying to enhance the present work using methods from previous work on sentiment lexicons for various languages and to incorporate crowdsourcing methods in order to further expand the data. It is also a limitation of the present work that it mostly focuses on single-word items, meaning that future work might focus on adding more multi-word expressions to the resource.

## Acknowledgements

We would like to thank other members of the Language and Technology lab at the University of Iceland for helpful discussions on this project. This work was supported by the Icelandic Student Innovation fund, grant nr. 2065110091.

## References

Arnardóttir, Þ. and Ingason, A. K. 2020. A neural parsing pipeline for Icelandic using the Berkeley Neural Parser. In Navarretta, C. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference 2020*, pages 48–51.

Burnap, P. and Williams, M. L. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242.

Çöltekin, Ç. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Friðriksdóttir, S. R. and Ingason, A. K. 2020. Disambiguating confusion sets in a language with rich morphology. In *Proceedings of ICAART 12*.

Ingason, A. K., Helgadóttir, S., Loftsson, H., and Rögnvaldsson, E. 2008. A mixed method lemmatization algorithm using a hierarchy of linguistic identities (HOLI). In Nordström, B. and Ranta, A., editors, *Advances in Natural Language Processing*, pages 205–216, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ingason, A. K., Jóhannsson, S. B., Rögnvaldsson, E., Loftsson, H., and Helgadóttir, S. 2009. Context-sensitive spelling correction and rich morphology. In Jokinen, K. and Bick, E., editors, *Proceedings of NoDaLiDa 2009*, pages 231–234.

Ingólfsdóttir, S. L., Loftsson, H., Daðason, J. F., and Bjarnadóttir, K. 2019. Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of NoDaLiDa 2019*.

Jökulsdóttir, T. F., Ingason, A. K., and Sigurðsson, E. F. 2019. A parsing pipeline for Icelandic based on the IcePaHC corpus. In Simov, K. and Eskevich, M., editors, *Proceedings of CLARIN Annual Conference 2019*, pages 138–141.

Kwok, I. and Wang, Y. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1621–1622.

Mollberg, D. E., Jónsson, Ó. H., Þorsteinsdóttir, S., Steingrímsson, S., Magnúsdóttir, E. H., and Gudnason, J. 2020. Samrómur: Crowd-sourcing data collection for Icelandic speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3463–3467.

Mubarak, H., Rashed, A., Darwish, K., Samih, Y., and Abdelali, A. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Nikulásdóttir, A. B., Guðnason, J., Ingason, A. K., Loftsson, H., Rögnvaldsson, E., Sigurðsson, E. F., and Steingrímsson, S. 2020. Language technology programme for Icelandic 2019-2023. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 3414–3422.

Þorsteinsson, V., Óladóttir, H., and Loftsson, H. 2019. A wide-coverage context-free grammar for Icelandic and an accompanying parsing system. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 1397–1404.

Pitenis, Z., Zampieri, M., and Ranasinghe, T. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 5113–5119.

Pitsilis, G. K., Ramampiaro, H., and Langseth, H. 2018. Detecting offensive language in tweets using deep learning. *Applied Intelligence*, 12:4730–4742.

Risch, J., Stoll, A., Ziegele, M., and Krestel, R. 2019. hpiDEDIS at GermEval 2019: Offensive language identification using a German BERT model. In *Proceedings of the 15th Conference on Natural Language Processing KONVENS*.

Sigurbergsson, G. I. and Derczynski, L. 2019. Offensive language and hate speech detection for Danish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3498–3508.

Wu, Z., Zheng, H., Wang, J., Su, W., and Fong, J. 2019. BNU-HKBU UIC NLP team 2 at SemEval-2019 task 6: Detecting offensive language using BERT model. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 551–555.