National Language Technology Infrastructure Initiative on CLARIN-IS


CLARIN and the Icelandic National Language Technology Programme

CLARIN-IS is a fairly new CLARIN member currently building a broad collection of national language resources for use in language technology. As a CLARIN C-centre, CLARIN-IS is hosting metadata for various text and speech corpora, lexical resources, software packages and models. The providers of the resources are universities, institutions and private companies working on a national Icelandic LT infrastructure initiative.

Core Projects – Language Resources

- Text corpora
- Parallel text corpora
- Lexicographic resources
- Speech corpora – crowdsourced
- Speech corpora – high quality recordings

Core Projects – NLP tools

- Tokeniser
- Part-of-speech taggers
- Parsers
- Lemmatiser
- Named Entity Recogniser
- Semantic Analysis

Core Projects – Infrastructure Software

- Speech recognition
- Speech synthesis
- Machine Translation
- Spell and grammar checking

For all infrastructure projects, the aim is to publish open environments for adaption and further research and development, as well as to deliver software packages that can be integrated into other products. The delivered software will be designed for defined domains, with instructions for adaption to other domains.

Standards and Deliverables

Aim is to publish the majority of the language resources under the CC BY 4.0 license. All delivered software will be published under the Apache 2.0 license, or a comparable license. Data and software development guidelines and quality standards are coordinated across all projects. All deliverables will be stored and distributed by CLARIN.

Status August 2021

47 repositories have been published on CLARIN-IS. This number will increase substantially by the end of September / beginning of October, when the deliverables from the second project year will be published. The 47 repositories can be divided into text corpora (19), lexical resources (7), NLP-tools (7), machine translation (4), spell and grammar checking (8), speech corpora (2), speech models and related modules (6).

See: https://clarin.is/