

# Towards better structured online data with the project “News, opinions or something else? Modeling text varieties in the multilingual Internet”

Veronika Laippala<sup>1</sup>, Saara Hellström<sup>1</sup>, Sampo Pyysalo<sup>1</sup>, Liina Repo<sup>1</sup>, Samuel Rönqvist<sup>1</sup>, Anna Salmela<sup>1</sup> and Valteri Skantsi<sup>1,2</sup>

<sup>1</sup>TurkuNLP Group, <sup>1</sup>University of Turku, <sup>2</sup>University of Oulu

## Motivation

- Potential of online data restricted by lack of metadata
- No reliable information on text registers, such as *news* and *blog posts* (Biber & Conrad 2009)
- Lack of register information
  - Can lead to wrong conclusions (Koplenig 2017)
  - Impacts Natural Language Processing (Tiedemann et al. 2016).

## Methods

**BERT** (Bidirectional Encoder Representations from Transformers; (Devlin et al. 2018))

- Deep learning model pre-trained on a large corpus of unlabelled text
- Uses the Transformer encoder to analyse text from both directions -> deeper sense of language context
- Adapted to a multilabel setting, Finnish model by Virtanen et al. (2019)

**CNN** (Convolutional neural network)

- Deep neural networks that use convolution filters that scan text represented as matrix of word vectors, breaks it into important features and judges whether each feature matches the relevant label (Kim 2014)
- Applied here in a multilingual setting with MUSE vectors (Conneau et al. 2018; Laippala et al. 2019)

## Objectives

- Characterize the full range of registers found on the Internet
- Automatically identify registers from web-based language resources
- Main focus on six languages:
  - English, Finnish, Swedish, French + Spanish and German to come
  - Extension to dozens of languages in the future

## Data

- Raw web data from Finnish Internet Parsebank, Common Crawl and Universal Parsebanks (Zeman et al. 2017)
- > 100 billion words in 64 languages!
- Ongoing manual register annotation of samples in Finnish, Swedish, French
- English register annotations from *Corpus of Online Registers of English* (CORE; Egbert et al. 2015)

## Register categories

- Taxonomy created for the English CORE
- Seems to suit other languages → very minor modifications needed

**Narrative:** News reports/News blogs, Sports reports, Personal blog, Historical article, Short story / Fiction, Travel blog, Community blog, Online article

**Informational Description:** Description of a thing, Encyclopedia articles, Research articles, Description of a person, Information blogs, FAQs, Course materials, Legal terms / conditions, Report

**Opinion:** Reviews, Personal opinion blogs, Religious blogs/sermons, Advice

**Interactive discussion:** Discussion forums, Question-Answer forums

**How-to/instructional:** How-to/instructions, Recipes

**Informational persuasion:** Description with intent to sell, News+Opinion blogs/Editorials

**Lyrical:** Songs, Poems

**Spoken:** Interviews, Formal speeches, TV transcripts

## Encyclopedia Article in the annotation tool

The screenshot shows a web page with a text snippet in Swedish. A dropdown menu is open, listing various register categories. The 'Encyclopedia articles' category is selected. Other categories include Narrative general, News reports / news blogs, Sports reports, Personal blog, Historical article, Fiction, Travel blog, Community blogs, Online article, Informational Description (or Explanation), Research articles, Description of a person, Information blogs, FAQs, and Course materials. There are also checkboxes for 'Epävarma', 'Vierasta kieltä', 'Epätyypillinen tekstilajissaan', 'Paljon kommentteja', 'Erikoismerkkejä', 'Sisältää erillisiä tekstejä', 'Osa tekstistä puuttuu', 'Generoitu tekstiä', and 'Swe-meta'.

## Preliminary register identification results

Languages Train → Test	Training data (docs.)	No of classes	Setting + Model	PR-AUC*
English → English	33 916	53	Multilabel BERT (English)	77%
Finnish → Finnish	7528	38	Multilabel BERT (Finnish)	85,4%
En + FI → Finnish	19 002	6	Multiclass CNN	85.3%

\* AUC (area under the receiver operating characteristic curve) provides an aggregate measure of performance across all possible classification thresholds

## References

- Biber & Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.
- Conneau, Lample, Ranzato, Denoyer & Jégou. 2018. Word translation without parallel data. *ICLR 2018*. CoRR, abs/1710.04087.
- Devlin, Chang, Lee & Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL 2019* (2018). arXiv preprint arXiv:1810.04805.
- Egbert, Biber & Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.
- Kim. 2014. Convolutional neural networks for sentence classification. *Proceedings of EMNLP 2014*.
- Koplenig. 2017. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities*, 32(1):169–188.
- Laippala, Kyllönen, Egbert, Biber & Pyysalo. 2019. Toward multilingual identification of online registers. *Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 292-297.
- Zeman et al. 2017. CoNLL 2017 Shared Task: Multilingual parsing from raw text to universal dependencies. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–19.
- Virtanen, Kanerva, Ilo, Luoma, Luotolahti, Salakoski, Ginter & Pyysalo. 2019. *Multilingual is not enough: BERT for Finnish*. ArXiv preprint at arXiv:1912.07076