# The Current State of Finnish NLP

**Mika Hämäläinen**
Faculty of Arts
University of Helsinki
and Rootroo Ltd
mika.hamalainen@helsinki.fi

**Khalid Alnajjar**
Faculty of Arts
University of Helsinki
and Rootroo Ltd
khalid.alnajjar@helsinki.fi

## Abstract

There are a lot of tools and resources available for processing Finnish. In this paper, we survey recent papers focusing on Finnish NLP related to many different subcategories of NLP such as parsing, generation, semantics and speech. NLP research is conducted in many different research groups in Finland, and it is frequently the case that NLP tools and models resulting from academic research are made available for others to use on platforms such as Github.

### Tiivistelmä

Suomen kielen koneelliseen käsittelyyn on tarjolla paljon valmiita työkaluja ja resursseja. Tässä artikkelissa tarkastelemme viimeaikoina julkaistuja tieteellisiä artikkeleita, joissa keskitytään suomen kielen kieliteknologiaan. Tarkastelemme kieliteknologian eri alaluokkia, kuten jäsentämistä, tuottamista, semantiikkaa ja puheetta. kieliteknologista tutkimusta tehdään Suomessa monissa eri tutkimusryhmissä, ja usein akateemisen tutkimuksen tuloksena tuotetut kieliteknologian työkalut ja mallit julkaistaan muiden käytettäväksi esimerkiksi Githubissa.

## 1 Introduction

There is no doubt that, within the Uralic language family, Finnish is one of the most well-resourced languages in terms of natural language processing (NLP). This has, however, not always been the case. Currently, NLP research conducted for Finnish has started to fragment into research outputs of several different research groups, and there is no survey paper out there that would describe the current state of Finnish NLP.

We hope that this survey paper clarifies the current situation and makes it clearer for people working in the academia outside of Finnish universities or in the industry and also for students. As it has been discussed before (Hämäläinen, 2021), Finnish is certainly not a low-resourced language, and our current survey further proves this point.

It is also important for researchers working on other smaller Uralic languages to see what has been done for Finnish in terms of NLP to see what the possible and meaningful directions are for further developing the resources needed. Especially since Uralic language share the same feature of rich morphology, which is something that commonly causes problems for computers.

## 2 Finnish NLP

In this section, we present a survey on the current state of Finnish NLP. We have tried to gather most of the current research on the topic, but we are certain that there are some research out there we have not been able to find. We have categorized the surveyed research outputs into parsing, generation, semantics and speech.

### 2.1 Parsing

Starting from morphology, stemming and spell checking Finnish is well supported in multiple commercial applications such as Microsoft and Google products. In the open-source world, low-level tasks such as stemming and spell checking can be conducted with Voikko[1].

Omorfi (Pirinen, 2015)[2] is currently the most well supported tool for morphological analysis (in-

[1]https://voikko.puimula.org/
[2]https://github.com/flammie/omorfi

cluding lemmatization) and generation. It is an FST (finite-state transducer) based tool developed on HFST (Helsinki finite-state technology) (Lindén et al., 2013) and it works together with constraint grammar (CG) based disambiguators and syntactic parsers available in the Giellatekno (Moshagen et al., 2014) repositories[3].

FinnPos[4] (Silfverberg et al., 2016) is another morphological tagger and lemmatizer tool based on CRF (conditional random field). There have been recently more data driven approaches focusing on Finnish (Silfverberg and Hulden, 2018).

While rule-based tradition has been strong in the past[5], there are several machine learning driven dependency parsers for Finnish, such as the statistical one[6] (Haverinen et al., 2014) and neural one[7] (Kanerva et al., 2018) by TurkuNLP.

Out of the aforementioned tools Omorfi (and the CG disambigator) and the machine learning based parsers are available to use through a Python package named UralicNLP[8] [9] (Hämäläinen, 2019).

As Finnish data is available in several multilingual datasets, there are many multilingual approaches for parsing (Qi et al., 2020)[10] (Honnibal et al., 2020)[11] and morphology (Aharoni and Goldberg, 2017; Nicolai and Yarowsky, 2019; Silfverberg and Tyers, 2019; Grönroos et al., 2020).

The fact that spoken Finnish is very different to standard Finnish has drawn some attention in the past (Jauhiainen, 2001) and recently (Partanen et al., 2019). The latter leading to a Python library called Murre[12] for automatic normalization of dialectal Finnish.

Non-standard data has been an issue in digital humanities (DH) projects (Mäkelä et al., 2020), and lately there have been efforts in automatically correcting OCR errors in existing historical datasets (Kettunen, 2015; Drobac and Lindén, 2020; Drobac, 2020; Duong et al., 2020).

Named entity recognition has also been under study with FiNER[13] and its recently released data (Ruokolainen et al., 2019). There is also another recent BERT (Devlin et al., 2019) based approach[14] to the topic (Luoma et al., 2020).

There have been several approaches to language detection including detection of Finnish from web corpora (see Jauhiainen et al., 2021). Similarly, native Finnish has been automatically identified from learner's Finnish (Malmasi and Dras, 2014).

In summary, parsing has been researched on different levels of language such as syntax, morphology, POS and NER tagging, and lemmatization. It has been mainly focusing on standard well-formed Finnish, although there are methods for coping with dialectal Finnish and OCR errors as well.

### 2.2 Generation

The lowest level of natural language generation is surface realization (see Reiter, 1994), and for that there are tools such as Omorfi and Syntax Maker[15] (Hämäläinen and Rueter, 2018). The latter uses Omorfi for morphological inflection while it takes care of higher level morphosyntax such as case government and agreement.

There is a strong computational creativity focus in Helsinki and it also shows in Finnish NLG, as there are several poem generators such as Keinoleino[16] (Hämäläinen, 2018b), Poeticus (Toivanen et al., 2012) and others (Hämäläinen and Alnajjar, 2019a,b). There is also an interactive poem generator tool called *Runokone* (Poem Machine)[17] (Hämäläinen, 2018c).

Recently there have been several approaches to enhancing existing news headlines (Alnajjar et al., 2019; Rämö and Leppänen, 2021). And some approaches to generating entire news articles automatically (Kanerva et al., 2019; Haapanen and Leppänen, 2020).

Paraphrase generation (Sjöblom et al., 2020) has also become a researched topic with the availability of monolingually aligned parallel corpora (Creutz, 2018). There is also an approach to converting standard Finnish text into different dialects (Hämäläinen et al., 2020).

Finnish is a typical language for machine translation tasks and it is not uncommon to see it featured in several papers that deal with multiple languages. However, there are several papers that fo-

[3]https://github.com/giellalt/lang-fin/tree/main/src/cg3
[4]https://github.com/mpsilfve/FinnPos
[5]See Pirinen, 2019b for some comparison between rules and neural networks
[6]https://turkunlp.org/Finnish-dep-parser/
[7]http://turkunlp.org/Turku-neural-parser-pipeline/
[8]https://github.com/mikahama/uralicNLP
[9]https://github.com/mikahama/uralicNLP/wiki/Dependency-parsing
[10]https://stanfordnlp.github.io/stanza/
[11]https://spacy.io/
[12]https://github.com/mikahama/murre
[13]https://github.com/Traubert/FiNer-rules/blob/master/finer-readme.md
[14]https://turkunlp.org/fin-ner.html
[15]https://github.com/mikahama/syntaxmaker
[16]https://github.com/mikahama/keinoleino
[17]http://runokone.cs.helsinki.fi/

cus on Finnish in particular (Hurskainen and Tiedemann, 2017; Hämäläinen and Alnajjar, 2019c; Pirinen, 2019a; Tiedemann et al., 2020).

There is also a recent approach to dialog generation in Finnish (Leino et al., 2020). Also non-native language learner's errors have been corrected successfully automatically (Creutz and Sjöblom, 2019).

To summarize the approaches, there are several generators for poetry and news that benefit from the available surface realizers. Paraphrasing, dialect adaptation, dialog generation and learners' error correction are domains with some research with potential for new discoveries in the future. Machine translation gets frequently attention from different researchers. There are several more NLG tasks (see Gatt and Krahmer 2018) that have not been researched at all in Finnish, which means that there is a lot of room for more research on this topic.

### 2.3 Semantics

Vector representations of meaning have become common place in NLP and Finnish is no exception with the availability of pretrained word2vec[18] [19] (Laippala and Ginter, 2014; Kutuzov et al., 2017) and fastText[20] (Bojanowski et al., 2017) models.

BERT models have also become available as part of the multilingual BERT model[21] (Devlin et al., 2019) or trained separately for Finnish[22] [23] (Kutuzov et al., 2017; Virtanen et al., 2019). Even Elmo models have been made available for Finnish[24] (Ulčar and Robnik-Šikonja, 2020).

In addition to the standard vector-based representations of meaning, there is another statistical model called SemFi[25] (Hämäläinen, 2018a). The model is a relational database that captures semantic relations of words based on their syntactic co-occurencies.

Before the era of machine learning, there were two prominent projects for modeling meaning computationally which have been translated into Finnish WordNet (Lindén and Carlson, 2010) and FrameNet (Lindén et al., 2019).

With the similar ideology to the hand crafted resources, there have been several different linked data projects in Finland representing semantics in structured ontologies (Hyvönen et al., 2006; Nyrkkö, 2018; Thomas et al., 2018; Koho et al., 2019). Many of the linked data projects are available on the Linked Data Finland website[26].

There is a Python library called FinMeter[27] (Hämäläinen and Alnajjar, 2019b) that has some higher level semantic tools for Finnish such as metaphor interpretation, word concreteness analysis and sentiment analysis. Sentiment analysis for Finnish has also been studied later on[28] (Öhman et al., 2020; Vankka et al., 2019; Lindén et al., 2020). There is also research on topic modeling methods (Ginter et al., 2009; Hengchen et al., 2018; Loukasmäki and Makkonen, 2019).

Finnish is well supported by traditional representations of semantics and latest vector based models. There is a vast amount of linked data resources in a variety of domains. Higher-level semantics such as metaphor interpretation and sentiment analysis also have received their share of research interest, although there are many more questions related to pragmatics and figurative language that have not been researched, such as sarcasm detection, multi-hop reasoning and fake news detection to name a few.

### 2.4 Speech

Apart from Finnish speech being supported by companies, there are some open-source tools that can synthesize Finnish. Festival[29] has a Finnish voice named Suopuhe[30], and eSpeak-ng[31] can even generate IPA characters for Finnish.

There are several more modern approaches to speech recognition (Enarvi et al., 2017; Varjokallio et al., 2021) and speech synthesis (Raitio et al., 2008, 2014). Although, speech synthesis has not gained much interest in the recent years.

There are several approaches to analyzing speech prosody (Virkkunen et al., 2018; Šimko et al., 2020). There is also some work on detecting different accents in spoken Finnish (Behravan et al., 2013, 2015) and named entity recognition (Porjazovski et al., 2020).

In summary, several approaches exist for speech processing in Finnish relating to recognition, ac-

[18] http://vectors.nlpl.eu/repository/
[19] https://bionlp.utu.fi/finnish-internet-parsebank.html
[20] https://fasttext.cc/docs/en/pretrained-vectors.html
[21] https://github.com/google-research/bert
[22] http://vectors.nlpl.eu/repository/
[23] https://github.com/TurkuNLP/FinBERT
[24] https://www.clarin.si/repository/xmlui/handle/11356/1277
[25] https://github.com/mikahama/uralicNLP/wiki/Semantics-(SemFi,-SemUr)
[26] https://www.ldf.fi/
[27] https://github.com/mikahama/finmeter
[28] a dataset https://github.com/Helsinki-NLP/XED
[29] https://www.cstr.ed.ac.uk/projects/festival/
[30] http://urn.fi/urn:nbn:fi:lb-20140730144
[31] https://github.com/espeak-ng/espeak-ng

cents and prosody. However, speech synthesis has received a surprisingly small amount of attention in the recent past. With the emergence of neural models, new research on synthesis could reach to potentially interesting new contributions.

## 3 Discussion and Conclusions

In this survey, we have gathered research conducted on different aspects of NLP. We have included links to models and code implementations for most of the research papers. It has been a pleasant thing to notice that not only Finnish NLP research exists but also it is often not conducted in a closed fashion, but the actual research outputs have been made openly available for a wider community of people even outside of academia. This is crucial for any language that is relatively small, like Finnish. If Finnish academics did not release their research, there would not be many other people in the world that would produce high-quality tools for Finnish.

Digital extinction is something that many endangered languages are facing right now (see Kornai 2013). Therefore, it is important to ensure that NLP resources become openly available for endangered Uralic languages as well. Availability itself is not enough, however, as the resources need to be easy to find and use. Despite the fact that we have open NLP tools for Finnish, we are still far a way from a world where machines use our language fluently. Finnair's in-flight entertainment system still announces happliy: **saavumme kohteeseen Helsinki* (*we arrive in destination Helsinki) instead of expressing it correctly, *saavumme Helsinkiin* (we arrive in Helsinki), Google Doc's spell checker does not recognize mostly any inflectional form with a possessive suffix and predictive text in mobile keyboards suggest overly formal normative Finnish only.

While Finnish NLP has come far in terms of academic research and tools built as a result, we as a nation are still far away from having Finnish language technology fully integrated into the systems we use every day. Many of the problems have been solved already, it is just the matter of the industry finding out about the NLP tools that are out there.

We have limited our survey to NLP tools and methods only. We know that there are a plethora of language resources available for Finnish as well. Based on our experiences, many corpora are well hidden and digging them up is a time consuming effort worthy of a separate survey paper. Unfortunately the Finnish practice of describing data on Metashare[32] is very unhelpful in this respect because the metadata descriptions in the service hardly ever contain information about where to access the data, how to cite it and who the real authors are.

## References


Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

Khalid Alnajjar, Leo Leppänen, Hannu Toivonen, et al. 2019. No time like the present: methods for generating colourful and factual multilingual news headlines. In *Proceedings of the 10th International Conference on Computational Creativity*. Association for Computational Creativity.

Hamid Behravan, Ville Hautamäki, and Tomi Kinnunen. 2013. Foreign accent detection from spoken finnish using i-vectors. In *INTERSPEECH*, volume 2013, page 14th.

Hamid Behravan, Ville Hautamäki, and Tomi Kinnunen. 2015. Factors affecting i-vector based foreign accent recognition: A case study in spoken finnish. *Speech Communication*, 66:118–129.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mathias Creutz and Eetu Eetu Sjöblom. 2019. Toward automatic improvement of language produced by non-native language learners. In *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019), September 30, Turku Finland*, 164, pages 20–30. Linköping University Electronic Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.


[32] https://metashare.csc.fi/

Senka Drobac. 2020. *OCR and post-correction of historical newspapers and journals*. Ph.D. thesis, University of Helsinki, Finland.

Senka Drobac and Krister Lindén. 2020. Optical character recognition with neural networks and post-correction with finite state methods. *International Journal on Document Analysis and Recognition (IJDAR)*, 23(4):279–295.

Quan Duong, Mika Hämäläinen, and Simon Hengchen. 2020. An unsupervised method for ocr post-correction and spelling normalisation for finnish. *arXiv preprint arXiv:2011.03502*.

Seppo Enarvi, Peter Smit, Sami Virpioja, and Mikko Kurimo. 2017. Automatic speech recognition with very large conversational finnish and estonian vocabularies. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(11):2085–2097.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Filip Ginter, Hanna Suominen, Sampo Pyysalo, and Tapio Salakoski. 2009. Combining hidden markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. *International journal of medical informatics*, 78(12):e1–e6.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2020. Morfessor em+ prune: Improved subword segmentation with expectation maximization and pruning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3944–3953.

Lauri Haapanen and Leo Leppänen. 2020. Recycling a genre for news automation: The production of valtteri the election bot. *AILA Review*, 33(1):67–85.

Mika Hämäläinen. 2018a. Extracting a semantic database with syntactic relations for finnish to boost resources for endangered uralic languages. *The Proceedings of Logic and Engineering of Natural Language Semantics 15 (LENLS15)*.

Mika Hämäläinen. 2018b. Harnessing nlg to create finnish poetry automatically. In *Proceedings of the ninth international conference on computational creativity*. Association for Computational Creativity (ACC).

Mika Hämäläinen. 2018c. Poem machine-a co-creative nlg web application for poem writing. In *The 11th International Conference on Natural Language Generation Proceedings of the Conference*. The Association for Computational Linguistics.

Mika Hämäläinen and Khalid Alnajjar. 2019a. Generating modern poetry automatically in finnish. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing Proceedings of the Conference*. The Association for Computational Linguistics.

Mika Hämäläinen and Khalid Alnajjar. 2019b. Let's face it. finnish poetry generation with aesthetics and framing. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 290–300.

Mika Hämäläinen and Khalid Alnajjar. 2019c. A template based approach for training nmt for low-resource uralic languages-a pilot with finnish. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 520–525.

Mika Hämäläinen, Niko Partanen, Khalid Alnajjar, Jack Rueter, and Thierry Poibeau. 2020. Automatic dialect adaptation in finnish and its effect on perceived creativity. In *11th International Conference on Computational Creativity (ICCC'20)*. Association for Computational Creativity.

Mika Hämäläinen and Jack Rueter. 2018. Development of an open source natural language generation tool for Finnish. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 51–58, Helsinki, Finland. Association for Computational Linguistics.

Katri Haverinen, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Anna Missilä, Stina Ojala, Tapio Salakoski, and Filip Ginter. 2014. Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48:493–531. Open access.

Simon Hengchen, Antti Olavi Kanner, Jani Pekka Marjanen, and Eetu Mäkelä. 2018. Comparing topic model stability between finnish, swedish, english and french. In *Digital Humanities in the Nordic Countries*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Arvi Hurskainen and Jörg Tiedemann. 2017. Rule-based machine translation from english to finnish. In *Proceedings of the Second Conference on Machine Translation*, pages 323–329.

Eero Hyvönen, Tuukka Ruotsalo, Thomas Häggström, Mirva Salminen, Miikka Junnila, Mikko Virkkilä, Mikko Haaramo, Eetu Mäkelä, Tomi Kauppinen, and Kim Viljanen. 2006. Culturesampo–finnish culture on the semantic web: The vision and first results. In *Developments in Artificial Intelligence and the Semantic Web-Proceedings of the 12th Finnish AI Conference STeP*, pages 26–27.

Mika Hämäläinen. 2019. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software*, 4(37):1345.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*. Rootroo Ltd.

Tommi Jauhiainen. 2001. Using existing written language analyzers in understanding natural spoken Finnish. In *Proceedings of the 13th Nordic Conference of Computational Linguistics (NODALIDA 2001)*, Uppsala, Sweden. Department of Linguistics, Uppsala University, Sweden.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2021. Suomalais-ugrilaiset kielet ja internet -projekti 2013-2019. In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*. Rootroo Ltd.

Jenna Kanerva, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakoski. 2018. Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.

Jenna Kanerva, Samuel Rönnqvist, Riina Kekki, Tapio Salakoski, and Filip Ginter. 2019. Template-free data-to-text generation of finnish sports news. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 242–252.

Kimmo Kettunen. 2015. Keep, change or delete? setting up a low resource ocr post-correction framework for a digitized old finnish newspaper collection. In *Italian Research Conference on Digital Libraries*, pages 95–103. Springer.

Mikko Koho, Esko Ikkala, Petri Leskinen, Minna Tamper, Jouni Tuominen, and Eero Hyvönen. 2019. Warsampo knowledge graph: Finland in the second world war as linked open data. *Semantic Web*, (Preprint):1–14.

András Kornai. 2013. Digital language death. *PloS one*, 8(10):e77056.

Andrei Kutuzov, Murhaf Fares, Stephan Oepen, and Erik Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 58th Conference on Simulation and Modelling*, pages 271–276. Linköping University Electronic Press.

Veronika Laippala and Filip Ginter. 2014. Syntactic n-gram collection from a large-scale corpus of internet finnish. In *Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT*, volume 268, page 184.

Katri Leino, Juho Leinonen, Mittul Singh, Sami Virpioja, and Mikko Kurimo. 2020. Finchat: Corpus and evaluation setup for finnish chat conversations on everyday topics. *Proc. Interspeech 2020*, pages 429–433.

Krister Lindén, Erik Axelson, Senka Drobac, Sam Hardwick, Juha Kuokkala, Jyrki Niemi, Tommi A Pirinen, and Miikka Silfverberg. 2013. HFST a system for creating NLP tools. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 53–71. Springer.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet–finnish wordnet by translation. *LexicoNordica–Nordic Journal of Lexicography*, 17:119–140.

Krister Lindén, Heidi Haltia, Antti Laine, Juha Luukkonen, Jussi Piitulainen, and Niina Väisänen. 2019. Finntransframe: translating frames in the finnframenet project. *Language Resources and Evaluation*, 53(1):141–171.

Krister Lindén, Tommi Jauhiainen, and Sam Hardwick. 2020. Finnsentiment–a finnish social media corpus for sentiment polarity annotation. *arXiv preprint arXiv:2012.02613*.

Petri Loukasmäki and Kimmo Makkonen. 2019. Eduskunnan täysistunnon puheenaiheet 1999-2014: miten käsitellä lda-aihemalleja? *Politiikka*, 61(2).

Jouni Luoma, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo. 2020. A broad-coverage corpus for finnish named entity recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4615–4624.

Eetu Mäkelä, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mika Hämäläinen, Samuli Kaislaniemi, Terttu Nevalainen, et al. 2020. Wrangling with non-standard data. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference Riga, Latvia, October 21-23, 2020*. CEUR-WS. org.

Shervin Malmasi and Mark Dras. 2014. Finnish native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2014*, pages 139–144.

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M. Tyers. 2014. Open-source infrastructures for collaborative work on under-resourced languages. The LREC 2014 Workshop "CCURL 2014 - Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era".

Garrett Nicolai and David Yarowsky. 2019. Learning morphosyntactic analyzers from the Bible via iterative annotation projection across 26 languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1765–1774, Florence, Italy. Association for Computational Linguistics.

Seppo Nyrkkö. 2018. Building a finnish som-based ontology concept tagger and harvester. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 18–25.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. Xed: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552.

Niko Partanen, Mika Hämäläinen, and Khalid Alnajjar. 2019. Dialect text normalization to normative standard Finnish. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 141–146, Hong Kong, China. Association for Computational Linguistics.

Tommi Pirinen. 2019a. Apertium-fin-eng–rule-based shallow machine translation for WMT 2019 shared task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 335–341, Florence, Italy. Association for Computational Linguistics.

Tommi A Pirinen. 2015. Development and use of computational morphology of finnish in the open source and open science era: Notes on experiences with omorfi development. *SKY Journal of Linguistics*, 28:381–393.

Tommi A Pirinen. 2019b. Neural and rule-based finnish nlp models—expectations, experiments and experiences. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 104–114.

Dejan Porjazovski, Juho Leinonen, and Mikko Kurimo. 2020. Named entity recognition for spoken finnish. In *Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery*, pages 25–29.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Tuomo Raitio, Heng Lu, John Kane, Antti Suni, Martti Vainio, Simon King, and Paavo Alku. 2014. Voice source modelling using deep neural networks for statistical parametric speech synthesis. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 2290–2294. IEEE.

Tuomo Raitio, Antti Suni, Martti Vainio, and Paavo Alku. 2008. Hmm-based finnish text-to-speech system utilizing glottal inverse filtering. In *Interspeech 2008, Brisbane, Australia, September 22-26, 2008*.

Miia Rämö and Leo Leppänen. 2021. Using contextual and cross-lingual word embeddings to improve variety in template-based nlg for automated journalism. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 62–70.

Ehud Reiter. 1994. Has a consensus nl generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation*.

Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2019. A finnish news corpus for named entity recognition. *Language Resources and Evaluation*, pages 1–26.

Miikka Silfverberg and Mans Hulden. 2018. Initial experiments in data-driven morphological analysis for Finnish. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 98–105, Helsinki, Finland. Association for Computational Linguistics.

Miikka Silfverberg, Teemu Ruokolainen, Krister Lindén, and Mikko Kurimo. 2016. Finnpos: an open-source morphological tagging and lemmatization toolkit for finnish. *Language Resources and Evaluation*, 50(4):863–878.

Miikka Silfverberg and Francis Tyers. 2019. Data-driven morphological analysis for uralic languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 1–14, Tartu, Estonia. Association for Computational Linguistics.

Eetu Sjöblom, Mathias Creutz, and Yves Scherrer. 2020. Paraphrase generation and evaluation on colloquial-style sentences. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1814–1822.

Suzanne Elizabeth Thomas, Anna Pia Frederike Wessman, Esko Ikkala, Jouni Antero Tuominen, Mikko Koho, Eero Antero Hyvönen, and Ville Rohiola. 2018. (co-) creating a sustainable platform for finland's archaeological chance finds: The story of sualt. In *Digital Heritage and Archaeology in Practice*. University press of Florida.

Jörg Tiedemann, Tommi Nieminen, Mikko Aulamo, Jenna Kanerva, Akseli Leino, Filip Ginter, and Niko Papula. 2020. The FISKMÖ project: Resources and tools for Finnish-Swedish machine translation and cross-linguistic research. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3808–3815, Marseille, France. European Language Resources Association.

Jukka Toivanen, Hannu Toivonen, Alessandro Valitutti, Oskar Gross, et al. 2012. Corpus-based generation of content and form in poetry. In *Proceedings of the third international conference on computational creativity*. University College Dublin.

Matej Ulčar and Marko Robnik-Šikonja. 2020. High quality elmo embeddings for seven less-resourced languages[33] In *Proceedings of The 12th Language*

[33]We don't agree with the title of the paper declaring Finnish as a "less-resourced" language. As we have seen in this paper Finnish does have a bunch of resources!

*Resources and Evaluation Conference*, pages 4731–4738.

Jouko Vankka, Heikki Myllykoski, Tuomas Peltonen, and Ken Riippa. 2019. Sentiment analysis of finnish customer reviews. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 344–350.

Matti Varjokallio, Sami Virpioja, and Mikko Kurimo. 2021. Morphologically motivated word classes for very large vocabulary speech recognition of finnish and estonian. *Computer Speech & Language*, 66:101141.

Päivi Johanna Virkkunen, Juraj Simko, Heini Henriikka Kallio, Martti Tapani Vainio, et al. 2018. Prosodic features of finnish compound words. In *Proceedings of the 9th International Conference on Speech Prosody 2018*. International Speech Communications Association.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Juraj Šimko, Martti Vainio, and Antti Suni. 2020. Analysis of speech prosody using WaveNet embeddings: The Lombard effect. In *Proc. 10th International Conference on Speech Prosody 2020*, pages 910–914.