Help from the Neighbors: Estonian Dialect Normalization Using a Finnish Dialect Generator

Mika Hämäläinen, Khalid Alnajjar, Tuuli Tuisk

▶ **To cite this version:**

Mika Hämäläinen, Khalid Alnajjar, Tuuli Tuisk. Help from the Neighbors: Estonian Dialect Normalization Using a Finnish Dialect Generator. Third Workshop on Deep Learning for Low-Resource Natural Language Processing, Jul 2022, Seattle, United States. hal-03718891

HAL Id: hal-03718891
https://hal.science/hal-03718891

Submitted on 9 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Help from the Neighbors: Estonian Dialect Normalization Using a Finnish Dialect Generator

Mika Hämäläinen[1], Khalid Alnajjar[1] and Tuuli Tuisk[2]
[1]École Normale Supérieure and CNRS, Paris, France
[2]University of Tartu, Estonia
firstname.lastname@{cnrs.fr[1],ut.ee[2]}

Abstract

While standard Estonian is not a low-resourced language, the different dialects of the language are under-resourced from the point of view of NLP, given that there are no vast hand normalized resources available for training a machine learning model to normalize dialectal Estonian to standard Estonian. In this paper, we crawl a small corpus of parallel dialectal Estonian - standard Estonian sentences. In addition, we take a savvy approach of generating more synthetic training data for the normalization task by using an existing dialect generator model built for Finnish to "dialectalize" standard Estonian sentences from the Universal Dependencies tree banks. Our BERT based normalization model achieves a word error rate that is 26.49 points lower when using both the synthetic data and Estonian data in comparison to training the model with only the available Estonian data. Our results suggest that synthetic data generated by a model trained on a more resourced related language can indeed boost the results for a less resourced language.

1 Introduction

Estonian itself can hardly be characterized as low-resourced due to a variety of NLP tools (Orasmaa et al., 2016; Kaalep et al., 2018; Laur et al., 2020) and corpora (Kaalep et al., 2010; Altrov and Pajupuu, 2012; Muischnek et al., 2016) available for the language. What still remains a difficult and severely under-resourced task to tackle is non-standard dialectal language. Estonian has a rich morphology which means that an individual word can have several different inflectional forms. In terms of dialects, this means that all of the different inflectional forms of a given word can be slightly different in different dialects of the language. This poses a challenge for NLP methods that are mostly trained on standard Estonian.

While the written standard is something people follow when they write official text such as published books or newspapers, people tend to communicate using dialect in more informal settings such as when sending messages or emails with friends and family or when engaging in discussion on online forums. This type of an every day language use is beyond the reach of current NLP methods for Estonian.

For other languages such as Finnish (Partanen et al., 2019), Swedish (Hämäläinen et al., 2020a) and German (Scherrer et al., 2019), dialect normalization has been seen as good way of dealing with the issue of non-standard language. If a model can normalize dialectal text to a standard norm, then all normative language NLP models can be applied on that data. Normalization has been shown to improve results in a variety of tasks such as parsing (van der Goot et al., 2020) and neologism retrieval (Säily et al., 2021).

Unfortunately, Estonian does not have vast dialectal resources available with aligned normalizations with dialectal sentences. For this reason, we establish a new methodology for producing synthetic dialectal Estonian - standard Estonian sentence pairs using a Finnish dialect generation model. The data and the models presented in this paper have been released openly on Zenodo[1].

Estonian dialects are traditionally divided into northern and southern groups, that differ on phonological, morphological as well as on lexical levels. According to the general Estonian dialect classification (Pajusalu et al., 2018), there are three main dialect groups. (1) The North Estonian dialect group consists of the Eastern, Insular, Central, and Western dialects. (2) The Northeastern Coastal dialect group consists of the Northeastern and the Coastal dialects. (3) The South Estonian group consists of the Mulgi, Tartu, and Võru dialects. In recent decades, the question of Seto has been debated. The distinction between Seto and Võru has been justified for instance on a syntactic level

[1]https://zenodo.org/record/6558469

Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, pages 61 - 66
July 14, 2022 ©2022 Association for Computational Linguistics

(Lindström et al., 2014). The dominant contact languages for Estonian dialects are Swedish, Russian, Latvian, and Votic. Finnish, Ingrian and Livonian have influenced somewhat less (Lindström et al., 2019).

2 Related work

There have been several different approaches to text normalization in the past (Bollmann, 2019). In this section, we will give a quick overview of the common approaches.

Dialectal text normalization has been tackled by using normalization rules and heuristics (Bollmann et al., 2011; Khan and Karim, 2012; Sidarenka et al., 2013). Later on, algorithmic approaches have been used for the task (Saloot et al., 2014; Rehan et al., 2018; Poolsukkho and Kongkachandra, 2018).

Very frequently, normalization is modeled as a character-level machine translation task. There are several research papers that use a statistical machine translation approach with a character level n-gram language model of varying lengths (Schlippe et al., 2010; De Clercq et al., 2013; Schlippe et al., 2013; Scherrer and Erjavec, 2013).

More recently, neural machine translation has been used on a character level for the normalization task (Bollmann and Søgaard, 2016; Ruzsics et al., 2019; Hämäläinen et al., 2019). The approaches consist typically of a bi-directional LSTM model and an attention mechanism. This approach has also been used with word2vec to extract and train an OCR post-correction model in an unsupervised way (Hämäläinen and Hengchen, 2019).

With the emergence of general purpose language models, many recent papers present work on using such models for text normalization. BERT (Muller et al., 2019; Plank et al., 2020), BART (Bucur et al., 2021) and RoBERTa (Kubal and Nagvenkar, 2021), for instance, have all been use lately to solve the task.

3 Dialect data

Since there is no dialectal corpus with standard normalizations available for Estonian, we have to crawl one relying on the accessible resources. The Institute of Estonian Language has released some dialectal dictionaries online[2]. Some of these contain example sentences in one of the dialects and their normalization in standard Estonian. In particular, we found that the dialectal dictionaries for Mulgi[3], Kihnu[4] and Hiiu[5] dialects had such aligned dialectal-standard Estonian sentence pairs.

We proceeded to crawl the aforementioned dictionaries. The dictionaries do not have an index of lemmas or any other means of browsing them apart from search queries. For this reason, we use the full text query the online dictionaries have to find occurrences of a given word in anywhere within the dictionary entries. We do this query for the 10,000 most frequent words[6] recorded in the Eesti kirjakeele sagedussõnastik (Kaalep and Muischnek, 2002) which is based on a relatively large 1 million word corpus. This crawling approach leads to the same texts being crawled multiple times, and for this reason, we remove all duplicates.

Some of the dialectal example sentences have additional annotation such as stress marked on top of the vowels. We clean the data of any additional marking and punctuations so that we are left with characters that are part of the Estonian alphabets. Furthermore, we ensure that the dialectal sentence and its normalization have an equal number of words. This step is needed because sometimes the example sentences were not normalized or were not fully normalized. This way we can clear all wrongly aligned sentences from the data. This resulted in 14510 aligned dialectal-normative Estonian sentences.

In Table 1, we can see examples of the data. As we can see, sometimes the dictionary authors had adapted a very strict normalization strategy; on top of just normalizing the sentence to follow the standard Estonian morphology and orthography, they had occasionally normalized dialectal words to completely different words that are part of the standard language. This is different from the vast dialect corpus available for Finnish (Kotimaisten kielten keskus, 2014), where the normalization does not replace any existing words with different ones. This fact alone makes this Estonian corpus more difficult to normalize automatically.

We split the corpus randomly to 70% training, 15% validation and 15% testing. This split is used for all the models we train that include Estonian data in their training. All models are evaluated with this test split.

[2]https://portaal.eki.ee/sonaraamatud.html

[3]https://eki.ee/dict/mulgisuur

[4]http://www.eki.ee/dict/kihnu

[5]http://www.eki.ee/dict/hiiu

[6]https://www.cl.ut.ee/ressursid/sagedused/table1.txt

Dialectal	Normalized	Translation
na joove kõrdamisi ütest laasist	nad joovad kordamööda ühest klaasist	they take turns drinking from one glass
Siis oli tiädmätä jäen ning oksõndan	Siis oli teadvuse kaotanud ja oksendanud	Then he had lost consciousness and vomited
ärä tettä alatude inemistege tegemist	ära tee alatute inimestega tegemist	don't deal with naughty people

Table 1: Examples of the corpus

4 Dialect normalization

We train BERT-based (Devlin et al., 2018) models to do Estonian normalization using Transformers Python library (Wolf et al., 2020). We model the task as a sequence to sequence task, where the model is trained to predict a normalized version of a sentence given a dialectal sentence. The model consists of a BERT based encoder and decoder models similarly to the architecture proposed in Rothe et al. (2020).

We build our models on EstBERT[7] (Tanvir et al., 2021) which is a BERT model trained solely on Estonian data using the Estonian National Corpus. We train three models: one with Estonian only data, one only with synthetic data and one with both types of data. We train the models for 3 epochs.

4.1 Generating synthetic Finnish data

Because Finnish and Estonian are closely related languages, we want to experiment whether synthetically produced dialectal Finnish data can improve the Estonian normalization models. Standard Estonian is closer to dialectal Finnish than standard Finnish, so it makes sense that a Finnish dialect like data could improve the results. It is important to note at this stage that this is not a Finnish to Estonian translation task. Finnish and Estonian are two very different languages and a model that can translate between the two languages has hardly anything to do with dialect normalization.

We use the Finnish dialect generation models presented by Hämäläinen et al. (2020b) to convert standard Estonian sentences into a pseudo Estonian dialect. The dialect generation models are available through Murre Python library[8]. The dialect generator supports over 20 Finnish dialects, and we need to indicate which dialect we want to generate when we use the model. Ideally, we would like to pick the dialect closest to the Estonian dialectal data, because Finland is a relatively large country and dialects further away from Estonia are already linguistically rather distant.

In order to find out which Finnish dialect produces the most Estonian dialect like output, we generate a dialectal version for each standard Estonian sentence in our corpus in each Finnish dialect. We compare the WER (word error rate) of each dialectal output to the correct dialectal Estonian sentence in the corpus that corresponds to the normalized sentence that was used to produce the dialectal sentences.

The results of the experiment, as seen in Table 2, indicate that Etelä-Karjala dialect gives an output closest to the Estonian dialectal data. For this reason, we pick this dialect to adapt sentences from the Estonian Universal Dependencies (UD) treebanks to a pseudo Estonian dialect. The treebanks have some noise, so we filter out all sentences that contain alphabets that are not part of Estonian such as $\mathring{a}$, ϕ or ω because they are an indication of non-Estonian sentences or non-Estonian words appearing in a sentence. We want the correct Estonian data to be of a very high quality, so we ensure that only sentences that have correct Estonian alphabets are retained. We also clear the sentences from non-alphabets such as numbers, punctuations and emojis.

Estonian has slightly different vowels than Finnish. The same speech sound [y] is written *y* in Finnish and *ü* in Estonian. For this reason, we replace *ü* with *y* before we pass it to the dialect generation model, and then we replace *y*s back to *ü*s in the output. Estonian also has one more vowel Finnish does not have, *õ*. In practice, both Estonian *ö* and *õ* are mapped to a single vowel *ö* in the Finnish phonetic system. We deal with this by excluding all Estonian UD sentences that have both *ö* and *õ*, so that the input can have either *ö* or *õ*. In case the input has *õ*, it is first replaced with *ö* and after the dialectal form has been generated, all *ö*s are replaced back to *õ*s.

After the dialect adaptation, we do a simple post-processing where we match the voice of plosives of each word in the dialectal output and the standard Estonian input. This means that if the Estonian word contained voiced plosives *d, b* or *g* without their unvoiced variants and if the dialectal output

[7] tartuNLP/EstBERT

[8] https://github.com/mikahama/murre

Finnish dialect	WER
Etelä-Häme	0.84
Etelä-Karjala	0.80
Etelä-Pohjanmaa	0.83
Etelä-Satakunta	0.82
Etelä-Savo	0.83
Eteläinen Keski-Suomi	0.83
Inkerinsuomalaismurteet	0.81
Kaakkois-Häme	0.82
Kainuu	0.84
Keski-Karjala	0.82
Keski-Pohjanmaa	0.83
Länsi-Satakunta	0.81
Länsi-Uusimaa	0.81
Länsipohja	0.81
Läntinen Keski-Suomi	0.82
Peräpohjola	0.81
Pohjoinen Keski-Suomi	0.85
Pohjoinen Varsinais-Suomi	0.81
Pohjois-Häme	0.82
Pohjois-Karjala	0.84
Pohjois-Pohjanmaa	0.84
Pohjois-Satakunta	0.82
Pohjois-Savo	0.85

Table 2: The WER between the Finnish dialect generator output and the Estonian dialect sentence. The lower the WER, the closer the output is to Estonian dialect.

had the corresponding unvoiced variant *t, p* or *k*, we replace the unvoiced consonant with the voiced variant. For example, *lambad* (sheep) is dialectalized to *lampaat*, which we convert to *lambaad*. This is important because Finnish dialects often unvoice voiced consonants, whereas the Estonian ones use voiced plosives frequently.

The generated data consists of over 336000 synthetically generated samples where the source side is in pseudo Estonian dialect produced by the Finnish dialect generator for Etelä-Karjala dialect and the target is clean standard Estonian from the UD tree banks. We split this data into 85% for training and 15% for validation.

5 Results and evaluation

In this section, we present the results of our models using WERs. Word Error Rate[9] is a commonly used metric to assess the quality of normalization models as it shows how far away the normalization predicted by a computational model is from the ground truth in terms of substitutions, insertions and deletions. We also calculate a token level accuracy which shows how many times a token was correctly normalized in the exact position it appeared in the sentence.

[9]We use the implementation from https://github.com/nsmartinez/WERpp

	WER	Accuracy
No normalization	74.09	0.257
Estonian only	77.74	0.240
Synthetic data only	73.84	0.256
Synthetic data and Estonian data	**55.25**	**0.471**

Table 3: The results of the BERT model with different datasets

The results can be seen in Table 3. The first row of the table shows how far away the dialectal sentence is from the standard Estonian one without applying a normalization. The WER and the accuracy were the best for the model that was trained on both the synthetic data and the Estonian data. These results are far from perfect, as even the best model makes mistakes around half of the time. However, the results look promising in the sense that the data augmentation improved the results drastically. It is interesting to see that neither the synthetic data nor the Estonian data alone seem to take the model too far, but when combined the results are way better. This is probably due to the fact that the Estonian data is rather small and training a model solely based on it is difficult, and that the synthetic data, while it helps the model to learn a mapping from something that looks like Estonian to standard Estonian, is does not represent the true difference between real Estonian dialects and the standard language. It is to be said, that with the amount of data we have at hand, it is unlikely that the model can ever learn to normalize Estonian the same way the dictionary authors had normalized the dialectal sentences, because then the model would need to also learn a mapping between dialectal words and more standard language.

6 Conclusions

We have shown that despite Estonian not having enough data on its own to train a dialect normalization model, using a Finnish dialect generator model with some orthographic conversion rules to produce synthetic data can boost the results. Although the results were promising, the best WER is still relatively high. This is partially due to the normalization strategy used in the original data. Nevertheless we believe that experimenting more with synthetic data in the future can help us push the WER lower.

7 Acknowledgments

This work was supported in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute). The work was also supported by the CNRS funded International Research Network Cyclades (Corpora and Computational Linguistics for Digital Humanities).

References

Rene Altrov and Hille Pajupuu. 2012. Estonian emotional speech corpus: theoretical base and implementation. In *4th international workshop on corpora for research on emotion sentiment & social signals (ES3)*, pages 50–53. Citeseer.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. 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 3885–3898.

Marcel Bollmann, Florian Petran, and Stefanie Dipper. 2011. Rule-based normalization of historical texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 34–42.

Marcel Bollmann and Anders Søgaard. 2016. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 131–139, Osaka, Japan. The COLING 2016 Organizing Committee.

Ana-Maria Bucur, Adrian Cosma, and Liviu P. Dinu. 2021. Sequence-to-sequence lexical normalization with multilingual transformers. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 473–482, Online. Association for Computational Linguistics.

Orphée De Clercq, Bart Desmet, Sarah Schulz, Els Lefever, and Véronique Hoste. 2013. Normalization of dutch user-generated content. In *9th International conference on Recent Advances in Natural Language Processing (RANLP 2013)*, pages 179–188. Incoma.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mika Hämäläinen and Simon Hengchen. 2019. From the paft to the fiiture: a fully automatic nmt and word embeddings method for ocr post-correction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 431–436.

Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. 2020a. Normalization of different swedish dialects spoken in finland. In *GeoHumanities'20: Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, page 24–27, United States. ACM.

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

Mika Hämäläinen, Tanja Säily, Jack Rueter, Jörg Tiedemann, and Eetu Mäkelä. 2019. Revisiting nmt for normalization of early english letters. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. The Association for Computational Linguistics.

Heiki-Jaan Kaalep, Sjur Nørstebø Moshagen, and Trond Trosterud. 2018. Estonian morphology in the giella infrastructure. In *Baltic HLT*, pages 47–54.

Heiki-Jaan Kaalep and Kadri Muischnek. 2002. *Eesti kirjakeele sagedussõnastik*. Tartu Ülikool.

Heiki-Jaan Kaalep, Kadri Muischnek, Kristel Uiboaed, and Kaarel Veskis. 2010. The estonian reference corpus: Its composition and morphology-aware user interface. In *Baltic HLT*, pages 143–146.

Osama A Khan and Asim Karim. 2012. A rule-based model for normalization of sms text. In *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, volume 1, pages 634–641. IEEE.

Kotimaisten kielten keskus. 2014. Suomen kielen näytteitä, ladattava versio. Kielipankki.

Divesh Kubal and Apurva Nagvenkar. 2021. Multilingual sequence labeling approach to solve lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 457–464, Online. Association for Computational Linguistics.

Sven Laur, Siim Orasmaa, Dage Särg, and Paul Tammo. 2020. Estnltk 1.6: Remastered estonian nlp pipeline. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7152–7160.

Liina Lindström, Maarja-Liisa Pilvik, Mirjam Ruutma, and Kristel Uiboaed. 2019. On the use of perfect and pluperfect in estonian dialects: Frequency and language contacts. *Uralica Helsingiensia*, 1(14):155–193.

Liina Lindström, Kristel Uiboaed, and Virve-Anneli Vihman. 2014. Varieerumine tarvis-/vaja-konstruktsioonides keelekontaktide valguses. *Keel ja Kirjandus*, pages 8–9.

Kadri Muischnek, Kaili Müürisep, and Tiina Puolakainen. 2016. Estonian dependency treebank: from constraint grammar tagset to universal dependencies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1558–1565.

Benjamin Muller, Benoit Sagot, and Djamé Seddah. 2019. Enhancing BERT for lexical normalization. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 297–306, Hong Kong, China. Association for Computational Linguistics.

Siim Orasmaa, Timo Petmanson, Alexander Tkachenko, Sven Laur, and Heiki-Jaan Kaalep. 2016. EstNLTK - NLP toolkit for Estonian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2460–2466, Portorož, Slovenia. European Language Resources Association (ELRA).

Karl Pajusalu, Tiit Hennoste, Ellen Niit, Peeter Päll, and Jüri Viikberg. 2018. *Eesti murded ja kohanimed [Estonian dialects and place names]. 3rd edition*. Tallinn: Eesti Keele Sihtasutus.

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.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sanphet Poolsukkho and Rachada Kongkachandra. 2018. Text normalization on thai twitter messages using ipa similarity algorithm. In *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–5. IEEE.

Palak Rehan, Mukesh Kumar, and Sarbjeet Singh. 2018. A modular approach for social media text normalization. In *Information and Decision Sciences*, pages 187–195. Springer.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Tatyana Ruzsics, Massimo Lusetti, Anne Göhring, Tanja Samardžić, and Elisabeth Stark. 2019. Neural text normalization with adapted decoding and pos features. *Natural Language Engineering*, 25(5):585–605.

Tanja Säily, Eetu Mäkelä, and Mika Hämäläinen. 2021. From plenipotentiary to puddingless: Users and uses of new words in early english letters. In *Multilingual Facilitation*, pages 153–169. University of Helsinki.

Mohammad Arshi Saloot, Norisma Idris, and Rohana Mahmud. 2014. An architecture for malay tweet normalization. *Information Processing & Management*, 50(5):621–633.

Yves Scherrer and Tomaž Erjavec. 2013. Modernizing historical slovene words with character-based smt. In *BSNLP 2013-4th Biennial Workshop on Balto-Slavic Natural Language Processing*.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. Digitising swiss german: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53(4):735–769.

Tim Schlippe, Chenfei Zhu, Jan Gebhardt, and Tanja Schultz. 2010. Text normalization based on statistical machine translation and internet user support. In *Eleventh annual conference of the international speech communication association*.

Tim Schlippe, Chenfei Zhu, Daniel Lemcke, and Tanja Schultz. 2013. Statistical machine translation based text normalization with crowdsourcing. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8406–8410. IEEE.

Uladzimir Sidarenka, Tatjana Scheffler, and Manfred Stede. 2013. Rule-based normalization of german twitter messages. In *Proc. of the GSCL Workshop Verarbeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation*.

Hasan Tanvir, Claudia Kittask, Sandra Eiche, and Kairit Sirts. 2021. EstBERT: A pretrained language-specific BERT for Estonian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 11–19, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Rob van der Goot, Alan Ramponi, Tommaso Caselli, Michele Cafagna, and Lorenzo De Mattei. 2020. Norm it! lexical normalization for Italian and its downstream effects for dependency parsing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6272–6278, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.