Murreviikko – A Dialectologically Annotated and Normalized Dataset of Finnish Tweets

Olli Kuparinen
Department of Digital Humanities
University of Helsinki
olli.kuparinen@helsinki.fi

Abstract

This paper presents Murreviikko, a dataset of dialectal Finnish tweets which have been dialectologically annotated and manually normalized to a standard form. The dataset can be used as a test set for dialect identification and dialect-to-standard normalization, for instance. We evaluate the dataset on the normalization task, comparing an existing normalization model built on a spoken dialect corpus and three newly trained models with different architectures. We find that there are significant differences in normalization difficulty between the dialects, and that a character-level statistical machine translation model performs best on the Murreviikko tweet dataset.

1 Introduction

Dialectal variation is typical of user-generated content on social media, alongside other types of variation such as misspellings and emojis. Such language can be challenging for Natural Language Processing tools that are trained on standard language.

We present a dataset of dialectal Finnish tweets which have been manually annotated by dialect and normalized to standard Finnish spelling. The dataset can be used as a test set for further work in, for instance, dialect identification or dialect-to-standard normalization.

We further experiment with the latter, testing four different methods to normalize the tweets automatically: the publicly available RNN-based Murre normalizer (Partanen et al., 2019), a statistical machine translation system, a Transformer-based neural machine translation system, and a normalizer based on the pre-trained ByT5 model. To give an example of the task, the original dialectal text *oonko mää nähäny* should be replaced with the standard form *olenko minä nähnyt* ('have I seen').

The main contributions of the paper are:

- We collect a tweet dataset spanning three years.
- We manually annotate the dialects and normalize the tweets to be used in further work.
- We train three new normalization models on transcribed dialect data with different model architectures.
- We evaluate the normalization performance of our three models, as well as an existing normalization model, on the dataset.

2 Related Work

2.1 Collection of Dialectal Content from Social Media

There have been a lot of efforts in recent years to collect dialectal content from social media. Ljubešić et al. (2016) describe TweetGeo, a tool to collect data from Twitter with restrictions on geography, language and features. They use the tool to collect tweets from the language continuum of Bosnian, Croatian, Montenegrin, and Serbian. Likewise, Huang et al. (2016) collect tweets originating in the United States to study dialectal variation on social media.

Hovy and Purschke (2018) collect over 16 million Jodel posts from German-speaking areas and use the data for dialect clustering. Barnes et al. (2021) collect a dataset of Norwegian tweets and annotate them by language (Bokmål, Nynorsk, dialect, and mixed). The dataset is further annotated with POS tags in Mæhlum et al. (2022).

The MultiLexNorm (van der Goot et al., 2021) dataset includes data from social media in 12 languages or varieties and is collected mostly from Twitter. Even though the collection does not directly aim for dialectal content, it includes dialectal variation in addition to, for instance, orthographic variation.

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 31–39

	Tweets	Dialect	Standard	Swedish	English
2020	181	143	37	1	-
2021	203	142	55	3	3
2022	76	59	16	-	1
Total	460	344	108	4	4

Table 1: Distribution of the tweets by year and language. Dialect and standard refer to Finnish. Five dialectal tweets from 2020 were deemed abusive and were excluded from the dataset.

2.2 Normalization

Lexical normalization has been used especially in the domain of historical texts (e.g., Pettersson et al., 2014; Bollmann, 2019). The recent MultiLexNorm shared task addressed the normalization of a multilingual dataset of user-generated content (van der Goot et al., 2021), and some work has also been conducted on dialect normalization (Scherrer and Ljubešić, 2016; Abe et al., 2018; Partanen et al., 2019).

Methodologically, character-level statistical machine translation models have been proposed for normalization tasks (e.g., Pettersson et al., 2014; Scherrer and Ljubešić, 2016; Hämäläinen et al., 2018). More recently, neural machine translation models have been used, either based on recurrent networks with attention (e.g., Abe et al., 2018; Partanen et al., 2019), or on the Transformer architecture (Tang et al., 2018; Wu et al., 2021; Bawden et al., 2022). Finally, the best performance in the MultiLexNorm shared task (Samuel and Straka, 2021) was obtained by fine-tuning byT5, a byte-level pre-trained model (Xue et al., 2022).

3 Murreviikko

Murreviikko ('dialect week') is a Twitter campaign initiated at the University of Eastern Finland which aims to promote the use of dialects in Finland on social media. The campaign has run for three years (2020, 2021, 2022) and lasts for one week in October.

3.1 Data Collection from Twitter

We collected tweets that included the keyword *murreviikko* or *#murreviikko* via the Twitter API. Our data comes from all three years (2020–2022). The yearly and language-wise distribution of the tweets is presented in Table 1. Future augmentation of the dataset is possible if the campaign is continued.

3.2 Dialectal Annotation

The collected tweets were first annotated with the language they include (dialectal Finnish, standard Finnish, Swedish or English; see Table 1).[1] After this initial stage, the dialectal tweets were checked for abusive content and five such tweets were removed from the dataset, leaving 344 dialectal tweets in total.

The dialectal Finnish tweets were annotated on two levels: following the two-way division of Finnish dialects (Eastern–Western) and the seven-way division traditionally used in Finnish dialectology, based on Kettunen (1940). An eighth dialect area is often distinguished between South-West and Häme[2], called transitional Southwestern dialects. Since it shares many features with South-West and Häme, it would be hard to discern it from these in a single tweet. It is thus left out of this study. The dialect areas are presented in Figure 1.

The traditional division is based mostly on morphological and phonological features. The annotation of the tweets is based on these same features. The features include, for instance, several diphthong changes and different gemination cases, as well as case markers, elision, consonant gradation variation, and personal pronouns. For most cases the annotation is straightforward based on these features. Tweets that are not recognizable or include mixed features are deemed to their own class.

The traditional division does not account for the capital Helsinki due to its history as a Swedish-speaking city. There are however nine tweets written in Helsinki slang (a mainly Häme dialect with a wealth of Swedish loanwords). Another dialect group (Helsinki) was thus added to the annotation to accommodate these tweets.

Table 2 presents the dialectal distributions of the tweets, which mostly follows the population densities of the areas, except for the city of Helsinki, which is seriously underrepresented. The Savo dialect is also overrepresented, which might be explained by the fact that the University of Eastern Finland, where the campaign is initiated, is located in Savo and the official tweets of the campaign are written in that dialect.

[1]The annotation and normalization is performed by the author, who holds a PhD in Finnish with a special focus on language variation.

[2]Häme is sometimes referred to with its Swedish name Tavastia.

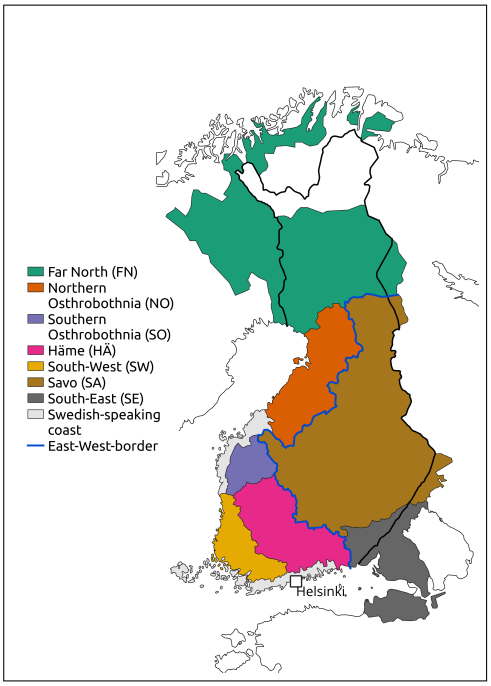

Figure 1: The seven dialect areas of Finnish, the East-West border (blue line) and the capital Helsinki. The dialect areas presented reflect the situation before World War II, when data was collected comprehensively (Kettunen, 1940). Modern-day dialects are mostly spoken inside the current borders of Finland, presented in black. The Northern Ostrobothnia in the map also includes Central Osthrobothnia which shares the dialect. The Northernmost areas are Sámi-speaking.

West	SW	74
	HÄ	58
	SO	17
	NO	33
	FN	14
	HE	9
	NA	12
	Total	217
East	SA	95
	SE	14
	NA	1
	Total	110
Unknown/Mixed		17

Table 2: Distribution of the tweets dialect-wise. The abbreviations are the same as in Figure 1 and HE=Helsinki. NA refers to tweets which contain dialectal language but are not distinguishable due to conflicting or scarce dialectal features. There might also be cases where the two-way division is distinguishable, but more fine-grained annotation is not possible.

3.3 Normalization

The dialectal tweets were manually normalized, following mostly the same principles as in the Samples of Spoken Finnish corpus (see Section 4.1). In essence, the tweets are normalized to a phonological and morphological standard, but word order is not altered, nor grammar rules of standard Finnish followed otherwise.

To give some examples of the phonological and morphological normalization, open or reduced diphthongs are returned to the standard alternative (*nuari* > *nuori* 'young', *koera* > *koira* 'dog'), weak grade alternatives of *t* are substituted with the standard *d* (*tehrä* > *tehdä* 'to do') and inessive case endings are presented with the standard *-ssa* or *-ssä* (*talos* > *talossa* 'in a house').

The principle has been to not distance the normalizations too far from the original dialects with insertions or word substitutions. An example of the principle is that possessive suffixes (*minun kirja**ni***, 'my book-my') are not added if they are not present in the original tweet (*mun kirja*, 'my book'), even though they are a part of standard Finnish. Likewise, dialect words are not corrected to the standard alternative, even if such words would exist, but instead normalized phonetically and morphologically (*seki diggaa fisuist* > *sekin diggaa fisuista* instead of *hänkin pitää kaloista* 's/he likes fish also').

The tweets include emojis, URLs, user mentions and hashtags. For the normalization experiments, emojis and URLs are removed from both the original and normalized side, user mentions are replaced with @@, and hashtags are normalized with the same rules as plain text.

The original text and normalization are aligned on tweet level. The dataset is accessible in compliance with the rules of the Twitter API, and the European Union's Digital Single Market directive (2019/790). This means that the tweet IDs, dialect annotations and corresponding normalizations are publicly available on Github.[3] The original tweets can be shared non-publicly for scientific use.

4 Normalization Experiments

4.1 Training Data

We use the Samples of Spoken Finnish (Institute for the Languages of Finland, 2021), hereafter SKN, for training. The corpus consists of 99 transcribed

[3]The public data is available at https://github.com/Helsinki-NLP/murreviikko. Licence: CC-BY-SA 4.0.

interviews from the 1960s that represent the dialects of Finnish comprehensively.[4] There are 50 Finnish-speaking locations in the corpus, with two speakers always representing a location (with one exception). The speakers are old and rural men and women, who have been born in the end of the 19th century (thus 70 to 90 years old at the time of the interview). The utterances of each interview have been randomly sampled and split to training (80%), development (10%) and test sets (10%).

The SKN corpus includes two transcription layers: one with very high precision, and a simplified version. Both rely on the Uralic Phonetic Alphabet (UPA), but the simplified transcriptions use almost exclusively standard Finnish characters and no diacritics. We use this version for training our own models. In contrast, the detailed transcriptions have been used to train the Murre normalizer (Partanen et al., 2019), which we will also experiment with. The transcriptions have been normalized to a phonetic standard manually by linguists. The principles of the normalization procedure are explained in the corpus, and they have been used as a guideline for the normalization of the tweet dataset (see Section 3.3).

Even though the simplified transcriptions use the same alphabet as the tweets, there are differences in, for instance, sandhi phenomena, which are marked in the transcriptions (*tehdäs se*) 'to do it', but often not in written dialectal Finnish (*tehdä se*). Likewise, the lexis used in old, rural interviews is naturally very different from the one used in the tweets. These are both issues that could affect the performance of the trained models.

Since the dialect transcriptions do not include any characters typical of social media, we add a set of 130 Finnish tweets to the training set. The tweets are collected from the OOD test set for Finnish Universal Dependencies[5], and added as such on both the original side and the normalized side. Such a small dataset makes the models aware of the special characters, but does not affect the normalization quality. The key figures of this dataset, along with those in the test set, are presented in Table 3.

	Sequences	Words	Words/Seq	Chars/Seq
Murreviikko	344	8269	24.04	175.25
SKN+UD$_{\text{tweets}}$	38,982	699,902	17.96	92.52

Table 3: Key figures of the datasets. Sequences refer to tweets on Murreviikko and UD$_{\text{tweets}}$ and utterances on SKN. Words/Seq = mean sequence length in words. Chars/Seq = mean sequence length in characters.

4.2 Methods and Tools

We treat normalization as a character transduction problem. This means that we split the sequences into individual characters and treat the characters as tokens, as has been standard practice in normalization tasks before (e.g., Scherrer and Ljubešić, 2016; Wu et al., 2021).

We experiment with four models:[6]

- **Murre.** The publicly available Murre normalizer[7] is based on a recurrent neural network (RNN) architecture and trained on the detailed transcriptions of the SKN corpus (Partanen et al., 2019). The Murre normalizer splits the data into non-overlapping trigrams and returns them to sentences in the output.

- **SMT.** Our statistical normalizer uses the Moses SMT toolkit (Koehn et al., 2007) with a character 10-gram KenLM language model trained on the training set. We do not use an additional language model on the target side. We use eflomal (Östling and Tiedemann, 2016) for character alignment. The model weights are tuned with minimum error rate training (MERT), with word error rate as the objective. Note that since we are working on characters, the word error rate is essentially character error rate.

- **NMT.** Our neural model follows standard Transformer architecture (Vaswani et al., 2017). It has 6 Transformer layers in the encoder and the decoder, with 8 heads each. There are 512 embedding and hidden layer dimensions. We use a batch size of 5000 tokens with an accumulate gradient of 4, and an initial learning rate of 4. The dropout is set to 0.1. We use position representation clipping with a value of 4 (Shaw et al., 2018). We train for 50,000 steps with checkpoints every 1000 steps. The model is trained with the OpenNMT-py toolkit (Klein et al., 2017).

[4] http://urn.fi/urn:nbn:fi:lb-2021112221, Licence: CC-BY.

[5] https://github.com/UniversalDependencies/UD_Finnish-OOD/, Licence: CC-BY-SA 4.0

[6] The training time and the number of parameters for each model are presented in Appendix A in Table 9.

[7] https://github.com/mikahama/murre, Licence: CC-BY-NC-ND 4.0

- **ByT5.** ByT5 (Xue et al., 2022) is a multilingual pre-trained sequence-to-sequence model which encodes all text as UTF-8 byte sequences (instead of subword tokenization), and uses the Transformer architecture. The model is pretrained on a masked language modeling task, where the model is asked to predict the content of a masked span. The data for pre-training is the multilingual m4C corpus (Xue et al., 2021), with 1.35% of the data being in Finnish. We use the `byt5-base` model and fine-tune it with our training data for 5 epochs, with maximum training sequence length of 512 bytes and a batch size of 4 sequences.

Our models are trained on sentence-level, whereas the tweets are left as they are and could thus include several sentences.

4.3 Evaluation

We evaluate the models on two metrics: character n-gram F-score (chrF2) and character error rate (CER). The former is typically used when evaluating machine translation models, and it calculates the F-score over character n-grams (Popović, 2015). CER is the Levenshtein distance between the model prediction and the correct target, normalized by the length of the target.[8]

We compare the systems to a **leave-as-is** (LAI) baseline, which evaluates the original sentences as they are, i.e., what would the scores be if the source was left untouched. For our own models, we also report the corresponding performance on a test set of the SKN corpus. This is not calculated with the Murre normalizer, since it is likely that some sentences in our test set were part of the training data for the model.

5 Results and Discussion

The chrF2 scores for the complete datasets are presented in Table 4. The statistical model performs best on the tweets (Murreviikko), with ByT5 achieving a very similar score. On the original dialect data (SKN) however, the best performance is obtained with the ByT5 model. The NMT model performs well on the original data, but does not generalize to the tweet dataset, as it barely outperforms the baseline. Likewise, the Murre normalizer does not produce a comparable score.

Model	Murreviikko	SKN
LAI	71.2	61.8
Murre	78.5	–
SMT	**84.4**	93.4
NMT	74.3	95.5
ByT5	83.6	**95.8**

Table 4: Character n-gram F-scores for complete datasets (↑).

Partanen et al. (2019)	5.73
SMT	7.95
NMT	5.32
ByT5	6.47

Table 5: Comparison of our models and Partanen et al. (2019) on the SKN corpus on word error rate (↓).

Partanen et al. (2019) present their results on the SKN dialect corpus on word error rate, which means the results presented in Table 4 are not directly comparable. To see how our models' performance relates to theirs, we present the word error rates of the models in Table 5, along with the score from Partanen et al. (2019). We calculated the word error rate with the same implementation as in the original work.[9]

Table 5 shows that our NMT model and the Murre normalizer (Partanen et al., 2019) offer very similar performance. The ByT5 model, which achieved the best chrF2 score, performs slightly worse when measured on word error rate. The models trained for this work are thus functioning on par with previous work for the dialect normalization task, but the performance does not translate to the tweet dataset.

To further analyze the difficulty of the tweet normalization task, we scrutinize the normalization performance on the different dialect groups to see if some dialects are inherently harder to normalize, or if some models fail on some dialects. The chrF2 scores broken down by dialect are presented in Table 6.

The baselines reflect that the South-Eastern (LAI 67.3) and especially South-Western dialects (LAI 59.3) are further from standard Finnish than the other dialects. Both dialects include for instance eli-

[8] We calculate chrF2 with the *sacrebleu* tool (Post, 2018), available at `https://github.com/mjpost/sacrebleu`, and CER with `https://github.com/nsmartinez/WERpp`.

[9] `https://github.com/nsmartinez/WERpp`.

Model	SW	HÄ	SO	NO	FN	HE	SA	SE	NA
LAI	59.3	71.7	73.1	74.9	75.6	73.9	74.2	67.3	83.7
Murre	75.1	78.8	79.1	81.1	77.2	70.4	80.4	78.3	79.6
SMT	**77.3**	85.4	**86.5**	87.5	**85.6**	73.2	**87.2**	**84.7**	88.8
NMT	62.9	75.8	76.7	78.1	77.1	73.6	77.8	68.9	83.5
ByT5	71.7	**85.9**	85.8	**87.6**	84.5	**83.4**	86.9	83.7	**91.0**

Table 6: Character n-gram F-scores dialect-wise (↑). SW = South-West, HÄ = Häme, SO = Southern Osthrobothnia, NO = Northern Osthrobothnia, FN = Far North, HE = Helsinki slang, SA = Savo, SE = South-East, NA = Not discernible.

Model	Murreviikko	SW	HÄ	SO	NO	FN	HE	SA	SE	NA	SKN
LAI	11.58	17.13	11.42	9.93	9.65	9.65	10.12	10.22	13.18	6.29	14.25
Murre	11.09	13.50	12.18	9.96	9.36	10.95	13.01	9.72	10.31	10.21	–
SMT	**7.64**	**11.13**	7.43	**6.91**	6.16	7.37	11.23	6.20	7.16	5.52	3.93
NMT	10.92	16.3	10.39	9.09	8.93	9.35	10.31	9.46	12.63	6.79	**1.84**
ByT5	7.72	13.49	**6.58**	11.06	**4.99**	**6.71**	**6.50**	**5.86**	**6.19**	**4.18**	2.37

Table 7: Character error rates for the complete datasets and dialect-wise. (↓).

sion and influence from other languages (Swedish and Estonian for the South-Western dialects, and other Finnic languages and Russian for the South-Eastern dialects). The rest of the dialect groups (disregarding NA) tend to have very similar baselines.

Regarding model performance, the Helsinki slang (HE) offers an interesting challenge. All models except ByT5 perform worse than the baseline. This is somewhat to be expected, as the training data does not include the slang. ByT5 on the other hand has been trained on web data by Common Crawl (Xue et al., 2021), which could include text written in the Helsinki slang. It could also be that the Swedish training data is helpful for the normalization task, since Helsinki slang is characterized by Swedish loanwords.

The difficulty of the South-Western dialects is reflected in the model scores, with all models achieving F-scores below 80. Given this is the second largest dialect group in the dataset, it also affects the overall performance quite significantly.

The character error rates for the complete datasets and dialects separately are presented in Table 7. The results follow mostly the same lines as the chrF2 scores presented in Table 4 and Table 6, but ByT5 achieves a better score on most dialects. However, it struggles with Southern Ostrobothnian and South-Western dialects so much that the statistical model achieves the best overall score on the whole dataset. Likewise for SKN, the NMT model performs better than ByT5 when evaluating on character error rate, whereas for chrF2 ByT5 achieved a better score.

5.1 Error Analysis

Table 8 presents an example sentence from a tweet with the predictions of each model. The example highlights common errors the models make. As South-Western dialects proved to be the hardest to normalize, the example is chosen from this dialect.

Murre fails to insert the hashtag and punctuation altogether. It has not seen the # in training (unlike our own models which were trained with the small tweet dataset added), and thus can not produce it. Likewise, it normalizes the *f* to *v* which is sometimes necessary in dialectal Finnish, but does not work well with the tweets which include a lot of loanwords from Swedish (such as the one in the example, *fundera* 'to think') and English.

However, Murre normalizes the morphological elements well, for instance managing to insert the correct adessive case ending *-lla* in *viikolla*, which is not achieved with any other model, as well as the ablative case ending *-ltä* in *sieltä*. Further fine-tuning of the model with modern text might thus produce comparable results.

The statistical model produces the hashtag and punctuation correctly, and also makes several correct substitutions and insertions (e.g., *päättys* >

Source	#Murreviikko päättys viime viikol, mut täsä muutmi fundeerauksi siält.
Target	#Murreviikko päättyi viime viikolla, mutta tässä muutamia fundeerauksia sieltä.
Murre	**M**urreviikko päättyi viime viikolla mutta tässä muu**tmin** **v**undeerauks**i** sieltä
SMT	#Murreviikko päättyi viime viiko**l**, mutta tässä muutam**i** fundeerauksia siel**t**.
NMT	#Murreviikko päättys viime viiko**l**, **mut** täsä muu**tmi** fundeerauks**i** sieltä
ByT5	#Murreviikko päättyi viime viiko**l**, mutta tässä muu**tmi** fundeerauks**i** siel**t**.
Gloss	'#Dialectweek ended last week, but here are some thoughts on it.'

Table 8: An example sentence from a tweet, with the source and correct target on top, and the corresponding normalizations of each model below. An English gloss is provided on the bottom. Errors of each model are presented in bold.

päättyi, mut > *mutta, täsä* > *tässä, fundeerauksi* > *fundeerauksia*), but fails to insert word-final characters in *viikol, muutami, sielt*.

The Transformer-based NMT consistently undernormalizes, producing predictions very close to the original source. The only difference in the example is the correctly normalized *siält* > *sieltä*. The prediction is also missing the final punctuation mark.

ByT5 has been originally trained on web crawled data, which enables the model to produce sensible output on the tweets. The errors are very similar to the ones produced by SMT, such as failing to insert word-final characters.

6 Conclusions

In this paper, we present a dataset of dialectal Finnish tweets which have been manually annotated by dialect and normalized to a standard form. The dataset will be made accessible to the scientific community for further testing and fine-tuning of models in the fields of dialect-to-standard normalization and dialect identification, for instance.

We furthermore evaluate four automatic normalization methods, which have been trained with transcribed spoken dialect data. Three of the models have been purpose-built for this paper, while a fourth model has been made publicly available (Partanen et al., 2019).

Character-level statistical machine translation provides the best normalization quality of the evaluated models on the Murreviikko-dataset, with the pre-trained and fine-tuned ByT5 model achieving very similar scores. Meanwhile, the ByT5 and a Transformer-based neural model perform best on the test set of the dialect transcriptions (SKN). The NMT model fails to transfer the performance to the tweets, however, consistently undernormalizing and barely outperforming the baseline. The RNN-based Murre normalizer struggles with the special characters typical of social media, while providing a reasonable performance on dialectal morphological features.

Dialect-wise, the South-Western dialects provide the lowest baseline and worst scores for the models. In the context of this work, it is thus the hardest to normalize from the traditional Finnish dialects. Helsinki slang, traditionally not seen as one of the dialects, is also difficult for the models but this is mostly due to a lack of training data.

Limitations

The size of the dataset is modest, and it is not possible to sensibly split it to train, development and test sets, for instance. We thus endorse it as a test set for future work.

We have not executed exhaustive hyperparameter tuning for our normalization experiments. It is likely that, for example, the neural machine translation model could perform better with further tuning and development. Likewise, we focus on character-level normalization and do not experiment with byte-pair encoding, found to enhance performance in recent normalization tasks (e.g., Bawden et al., 2022).

Acknowledgments

This work has been supported by the Academy of Finland through project No. 342859 "CorCoDial – Corpus-based computational dialectology".

References

Kaori Abe, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui. 2018. Multi-dialect neural machine translation and dialectometry. In *Proceedings of the*

32nd Pacific Asia Conference on Language, Information and Computation, Hong Kong. Association for Computational Linguistics.

Jeremy Barnes, Petter Mæhlum, and Samia Touileb. 2021. NorDial: A preliminary corpus of written Norwegian dialect use. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 445–451, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. 2022. Automatic normalisation of early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3354–3366, Marseille, France. European Language Resources Association.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Mika Hämäläinen, Tanja Säily, Jack Rueter, Jörg Tiedemann, and Eetu Mäkelä. 2018. Normalizing early English letters to present-day English spelling. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 87–96, Santa Fe, New Mexico. Association for Computational Linguistics.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding u.s. regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255.

Institute for the Languages of Finland. 2021. Samples of Spoken Finnish, VRT Version.

Lauri Kettunen. 1940. *Suomen murteet. 3, A, Murrekartasto*. Suomalaisen Kirjallisuuden Seuran toimituksia ; 188. Osa. Suomalaisen kirjallisuuden seura, Helsinki.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Nikola Ljubešić, Tanja Samardžić, and Curdin Derungs. 2016. TweetGeo - a tool for collecting, processing and analysing geo-encoded linguistic data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3412–3421, Osaka, Japan. The COLING 2016 Organizing Committee.

Petter Mæhlum, Andre Kåsen, Samia Touileb, and Jeremy Barnes. 2022. Annotating Norwegian language varieties on Twitter for part-of-speech. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 64–69, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

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.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2014. A multilingual evaluation of three spelling normalisation methods for historical text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 32–41, Gothenburg, Sweden. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

David Samuel and Milan Straka. 2021. ÚFAL at MultiLexNorm 2021: Improving multilingual lexical normalization by fine-tuning ByT5. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 483–492, Online. Association for Computational Linguistics.

Yves Scherrer and Nikola Ljubešić. 2016. Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS)*.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Gongbo Tang, Fabienne Cap, Eva Pettersson, and Joakim Nivre. 2018. An evaluation of neural machine translation models on historical spelling normalization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1320–1331, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Experimental Details

We trained the NMT model and ByT5 on a single NVIDIA V100 GPU. The CSMT model is trained on a Xeon Gold 6230 CPU. Table 9 presents the training time and number of parameters for the training data.

Model	Runtime (hh:mm)	Parameters
SMT	72:00	—
NMT	16:26	25.4 M
ByT5	9:56	581 M

Table 9: Training runtime and number of parameters for the training data.