MultiLexNorm: A Shared Task on Multilingual Lexical Normalization

Rob van der Goot[1], **Alan Ramponi**[2], **Arkaitz Zubiaga**[3], **Barbara Plank**[1], **Benjamin Muller**[4], **Iñaki San Vicente Roncal**[5], **Nikola Ljubešić**[6], **Özlem Çetinoğlu**[7], **Rahmad Mahendra**[8], **Talha Çolakoğlu**[9], **Timothy Baldwin**[10], **Tommaso Caselli**[11], **Wladimir Sidorenko**

[1]IT University of Copenhagen, [2]Fondazione Bruno Kessler, [3]Queen Mary University of London, [4]Inria Paris, [5]Elhuyar Foundation, [6]Jožef Stefan Institute, [7]University of Stuttgart, [8]Universitas Indonesia, [9]Huawei, [10]The University of Melbourne, [11]University of Groningen

robv@itu.dk

Abstract

Lexical normalization is the task of transforming an utterance into its standardized form. This task is beneficial for downstream analysis, as it provides a way to harmonize (often spontaneous) linguistic variation. Such variation is typical for social media on which information is shared in a multitude of ways, including diverse languages and code-switching. Since the seminal work of Han and Baldwin (2011) a decade ago, lexical normalization has attracted attention in English and multiple other languages. However, there exists a lack of a common benchmark for comparison of systems across languages with a homogeneous data and evaluation setup. The MULTILEXNORM shared task sets out to fill this gap. We provide the largest publicly available multilingual lexical normalization benchmark including 12 language variants. We propose a homogenized evaluation setup with both intrinsic and extrinsic evaluation. As extrinsic evaluation, we use dependency parsing and part-of-speech tagging with adapted evaluation metrics (a-LAS, a-UAS, and a-POS) to account for alignment discrepancies. The shared task hosted at W-NUT 2021 attracted 9 participants and 18 submissions. The results show that neural normalization systems outperform the previous state-of-the-art system by a large margin. Downstream parsing and part-of-speech tagging performance is positively affected but to varying degrees, with improvements of up to 1.72 a-LAS, 0.85 a-UAS, and 1.54 a-POS for the winning system.[1]

1 Introduction

The rise of social media has led to a tremendous increase in the amount of data shared over the Internet. But because of its spontaneous nature, the data naturally abounds with numerous language variations, both intended (e.g., slang, abbreviations, non-standard capitalization) and unintended ones (e.g., typos). This, in turn, poses considerable problems for existing natural language processing (NLP) tools (e.g., Baldwin et al., 2013; Eisenstein, 2013), most of which were originally designed to process canonical texts. One way to improve the performance of such systems is to *normalize* text and thus make it more similar to the data the NLP systems were initially designed for (and trained on).

At this point, to avoid confusion with other existing notions of text normalization (cf. Sproat et al., 2001; Aw et al., 2006), we should state that, throughout this paper, we will only deal with *lexical normalization*—a task which Han and Baldwin (2011) define as "a mapping from 'ill-formed' out-of-vocabulary (OOV) lexical items to their standard lexical forms." We focus only on social media data, as opposed to historical data (Tang et al., 2018; Bollmann, 2019) or medical data (Dirkson et al., 2019), and extend the scope of this task further to the cases where wrong in-vocabulary (IV) tokens can be normalized to (i.e., replaced with) their in-vocabulary counterparts, arriving at the following formulation:

Definition - Lexical Normalization

Lexical normalization is the task of transforming an utterance into its standard form, word by word, including both one-to-many (1-n) and many-to-one (n-1) replacements.

It should be noted that deletions and insertions of complete words are thus beyond the scope of the task as defined here.

Although lexical normalization potentially removes social signals (Nguyen et al., 2021), it has also been shown to boost many downstream NLP tasks, including named entity recognition (Schulz et al., 2016; Plank et al., 2020), POS tagging (Derczynski et al., 2013; Schulz et al., 2016; van der

[1]Data and submissions are available at https://bitbucket.org/robvanderg/multilexnorm/

Proceedings of the 2021 EMNLP Workshop W-NUT: The Seventh Workshop on Noisy User-generated Text, pages 493–509
November 11, 2021.

Lang.	Language name	Normalization example
DA	Danish	De skarpe lamper gjorde destromindre ek bedre . De skarpe lamper gjorde destro mindre ikke bedre .
DE	German	ogäj isch hätts auch dwiddern könn Okay ich hätte es auch twittern können
EN	English	u hve to let ppl decide what dey want to do you have to let people decide what they want to do
ES	Spanish	@username cuuxamee sii peroo veen yaa eem @username escúchame sí pero ven ya eh
HR	Croatian	svi frendovi mi nešto rade , veceras san osta sam . svi frendovi mi nešto rade , večeras sam ostao sam .
ID-EN	Indonesian-English	pdhal not fully bcs those ppl jg sih . padahal not fully because those people juga sih .
IT	Italian	a Roma è cosí primavera che sembra gia giov a Roma è così primavera che sembra già giovedì
NL	Dutch	Kga me wss trg rolle vant lachn Ik ga me waarschijnlijk terug rollen van het lachen
SL	Slovenian	jst bi tud najdu kovanec vreden veliko denarja . jaz bi tudi našel kovanec vreden veliko denarja .
SR	Serbian	komunalci kace pocne kaznjavanje ? komunalci kad počne kažnjavanje ?
TR	Turkish	He o dediyin suala cvb verdim He o dediğin suale cevap verdim
TR-DE	Turkish-German	@username Yerimm senii , damkee schatzymm :-* @username Yerim seni , danke Schatzym :-*

Table 1: Examples from MULTILEXNORM. One utterance per language: original sentence on top, and normalization on the bottom.

Goot et al., 2017; Zupan et al., 2019), dependency and constituency parsing (Baldwin and Li, 2015; van der Goot et al., 2020; van der Goot and van Noord, 2017), sentiment analysis (Van Hee et al., 2017; Sidarenka, 2019, pp. 79, 122), and machine translation (Bhat et al., 2018). However, existing work on this topic is largely fragmented, focused mostly on one language, relies on different evaluation metrics, or makes different assumptions regarding the items to be normalized (cf. Yang and Eisenstein, 2013; Li and Liu, 2015; Xu et al., 2015). All this makes it extremely hard to compare existing and new normalization systems.

In an attempt to achieve greater reproducibility, linguistic variety, and a standardized benchmark for *multilingual lexical normalization*, we introduce the MULTILEXNORM shared task. The benchmark for this task comprises datasets for 12 language(-pair)s: Danish, German, English, Spanish, Croatian, Indonesian-English, Italian, Dutch, Slovenian, Serbian, Turkish, and Turkish-German. All datasets contain sentences from popular social media platforms, which have been annotated for lexical normalization (i.e., with word-level replacements). Following our definition, we considered both intended and unintended spelling deviations, and included all categories defined by van der Goot et al. (2018) except phrasal abbreviations. We assume gold tokenization in all datasets, and leave automation of the tokenization step for future work. Examples of annotated sentences for all languages are shown in Table 1.

Furthermore, to precisely measure the effect of text normalization on downstream tasks, we also included a dedicated track on *extrinsic evaluation*, in which we estimate how much the results of dependency parsing and part-of-speech (POS) tagging change after normalization. This track includes corpora for English, German, Italian, and Turkish, annotated with Universal Dependencies (Nivre et al., 2020).

More details about intrinsic and extrinsic datasets are given in §2 and §5, respectively. We also provide an overview of baselines and submitted systems in §3, discussing their intrinsic and extrinsic results in §4 and §5, respectively. The paper concludes with a summary of the findings of the shared task and suggestions for future work.

Language and citation	Kappa	Same cand.
EN (Baldwin et al., 2015)	0.89	
EN (Pennell and Liu, 2014)	0.59	98.73
IT (van der Goot et al., 2020)	0.64–0.79	73.91–77.78
NL (Schuur, 2020)	0.77–0.91	
DA (Plank et al., 2020)	0.89	96.3

Table 2: Agreement scores for lexical normalization found in the literature. The "Kappa" column reports Fleiss/Cohen's kappa (rounded to 2 decimals) on the decision of whether a word needs to be normalized or not, whereas "Same cand." reports the raw percentage of times annotators agreed. Ranges in Italian include raw annotation and after some automatic fixes; for Dutch they are between different domains.

2 Data

Our selection of languages is purely based on dataset availability. We are aware that the benchmark contains mostly Indo-European languages, and encourage additions to this benchmark in the future to increase language variety.[2] We kindly request future work to cite the original data sources, and provide the bib-files on our website.

Lexical normalization is a subjective task, as in many cases multiple interpretations and annotations are plausible. Furthermore, annotators may disagree on what is "normal", and whether normalization is necessary for certain words. We have summarized all results on studies on inter-annotator agreement that we are aware of in Table 2.

Two types of agreement are reported in the literature: (1) Cohen's/Fleiss' kappa score on the choice of whether a word is in need of normalization; and (2) "Same candidate", which reports the percentage of times annotators agreed on the normalization replacement for words normalized by multiple annotators. Results in Table 2 show that the choice of whether to normalize has a medium to high kappa score, whereas the choice of the correct normalization candidate is generally high. An exception is Italian, which has a relatively low score due to some annotators not correcting capitalization (van der Goot et al., 2020).

Besides converting all data to the same format, we have attempted to converge annotation styles whenever possible. In particular, we applied the same normalization annotation in these cases:

- Interjections and punctuation are kept untouched, *hahaha* $\mapsto$ *hahaha* and not *haha*;
- Usernames, hashtags and URLs are kept untouched; if data is anonymized, usernames become @username;
- We kept capitalization correction where available. Unfortunately we did not have the budget to include capitalization correction in all datasets;
- We removed data that is not in the target language (mostly Frisian and Afrikaans in the Dutch data, and Indonesian and Dutch tweets in the German dataset);
- We fixed some tokenization issues in multiple languages.

With regard to data availability and composition, we note that some of the datasets were published before the shared task was held.[3] All datasets contain data from the Twitter platform; the Dutch corpus also includes forum and SMS data, and the Danish dataset includes texts from Arto, a Danish social media network. More details about the data collection for each dataset can be found in the dataset statement (Appendix A).

An overview of our datasets is shown in Table 3. It is clear that different annotation guidelines have been used, where some included "one-to-many" and "many-to-one" replacements, and correction of capitalization, where others did not. Furthermore, the amount of necessary normalization is very different, and the training splits of the datasets vary greatly, with the largest being almost 10 times larger than the smallest. It should also be noted that only 7 languages (DE, EN, HR, ID-EN, NL, SL, SR) have a dedicated development test, due to data availability.

3 Methods

3.1 Baselines

The organizers provided two naive baselines (i.e., LAI and MFR, as introduced below), and an "informed" baseline, based on training the previous state-of-the-art MoNoise over the respective datasets (van der Goot, 2019a).

[2] Please contact the first author of this paper if you are interested in adding a language. If more languages are added, future versions of MULTILEXNORM will be released via the repository.

[3] We removed them where possible. Specifically, 5 out of 12 languages still had their test data online when the competition started. However, they were not always easy to find, were in a different format, or the annotation differed from that used in the shared task.

Lang	Words	Sents	1-n	n-1	Change	Caps	Source
DA	16,448	719	0.29	0.04	9.25	+	(Plank et al., 2020)
DE	15,006	1,628	1.53	0.19	17.96	+	(Sidarenka et al., 2013)
EN	35,216	2,360	0.87	0.04	6.90	−	(Baldwin et al., 2015)
ES	7,189	568	0.00	0.00	7.69	−	(Alegria et al., 2013)
HR	54,416	4,760	0.01	0.00	8.89	−	(Ljubešić et al., 2017a)
ID-EN	13,949	495	1.43	0.19	12.16	−	(Barik et al., 2019)
IT	12,645	593	0.28	0.00	7.32	+	(van der Goot et al., 2020)
NL	12,381	907	5.98	0.06	28.29	+	(Schuur, 2020)
SL	44,944	4,670	0.00	0.00	15.62	−	(Erjavec et al., 2017)
SR	56,823	4,138	0.00	0.00	7.65	−	(Ljubešić et al., 2017b)
TR	6,443	570	3.00	1.69	37.02	+	(Çolakoğlu et al., 2019)
TR-DE	12,773	800	2.51	0.81	24.14	+	(van der Goot and Çetinoğlu, 2021)

Table 3: Some statistics on the 12 language(-pair)s within the MULTILEXNORM benchmark. The "1-n" column indicates the percentage of words which are split into multiple words (one-to-many), "n-1" indicates the proportion of words that are merged with other words as part of normalization (many-to-one), and "Change" indicates the percentage of words that are normalized. "Caps" indicates whether standard capitalization is included in the annotation; for datasets without annotation of capitalization, everything is lowercased.

LAI Leave-As-Is baseline, which simply returns the input word.

MFR Most-Frequent-Replacement baseline. It stores for every input word (unigram) its most frequent replacement in the training data. Then at run-time it simply replaces each word with its most common replacement. If a word is not seen before, it is returned as-is.

MoNoise This is based on a two-step approach. It first generates candidates based on: word embeddings, the Aspell spell checker,[4] replacements found in the training data, and some heuristics. In the second step, features from the generation step are combined with additional features (including character n-gram probabilities) and used to train a random forest classifier, which predicts the probability that a candidate is the correct candidate. The only tuned component is the generation of Aspell candidates, where the `--bad-spellers` options can be used to generate more candidates. For most languages this resulted in a slower but more effective model (except for HR, ID-EN, and SL). For the code-switched language pairs, the code-switched version of MoNoise was used (van der Goot and Çetinoğlu, 2021). To retrain MoNoise, new raw data was collected to base its n-gram probabilities and word embeddings on. We downloaded Twitter data of 2012–2020 from archive.org, filtered it with the fastText language classifier (Joulin et al., 2017a), and used the most recent Wikidump for each language.[5]

3.2 Submissions

The shared task ran in mid-2021, and attracted 9 participants with 18 submissions. We include the full list of submissions, but no system description or paper was received from **maet**, **team**, **thunderml**, or **learnML**, so the details of these methods are not clear. Submissions marked with an asterisk ("*") involve one or more of the shared task organizers.

ÚFAL (Samuel and Straka, 2021)

The system is based on ByT5 (Xue et al., 2021), and is a word-by-word normalization model. In order for the model to be as close to the original pre-training task as possible, each input word is normalized independently: it is enclosed in an opening and ending tag, over which ByT5 is run to produce the normalization.

The authors fine-tune ByT5 in two steps, first on synthetic data and then on the MULTILEXNORM data. To obtain synthetic data, they use Wikipedia as target data, and create unnormalized input through character edits, word edits, and dictionary replacements trained from the MULTILEXNORM data. During the final fine-tuning, they either: (a) use only the MULTILEXNORM data; or (b) com-

[4] http://aspell.net/

[5] Of 01-08-2021. Available at https://robvanderg.github.io/blog/twit_embeds.htm

bine the MULTILEXNORM with the synthetic data. They submitted two systems, a single model for every treebank and an ensemble of 4 models for every treebank. Both adapting the input to fit the pre-training step and the use synthetic data proved to be very beneficial for the system.

HEL-LJU* (Scherrer and Ljubešić, 2021)

The system is based on a BERT (Devlin et al., 2019) token classification preprocessing step, where for each token the type of the necessary transformation is predicted (none, uppercase, lowercase, titlecase, modify), and a character-level statistical machine translation (SMT) model is used to normalize accordingly. For some languages, depending on the results on the development data, the training data was extended by back-translating OpenSubtitles data. The paper evaluates a range of MT systems and ablations, and shows that a character-level SMT model is highly competitive.

TrinkaAI (Kubal and Nagvenkar, 2021)

The proposed model is based on a sequence labeling approach, where the input tokens are unnormalized and the target tokens are normalized. To reduce the target labels and make predictions faster, classes are based on those tokens for which normalization is required, and tokens which do not need to be normalized are labelled with a single target token. This sequence labeling model is fine-tuned on a pre-trained multilingual model encompassing all languages in the shared task. Further, a post-processing layer concerning word-alignment is applied, which further improved performance. This sequence-labeling approach ranked 6th out of 21 models, and scored highest among all competitors for the Spanish Language.

BLUE (Bucur et al., 2021)

The team tackled the task of lexical normalization as a neural machine translation problem, using the MBart-50 (Tang et al., 2020) multilingual many-to-many model. They fine-tuned the model for all the available languages, and used a MFR baseline for Danish and Serbian. They opted for a sentence-level approach as opposed to a word-level approach, using simple linear sum assignment based on Levenshtein distance to align the normalized words with the raw words.

CL-MoNoise* (van der Goot, 2021)

This is the same method as MoNoise, but it is deployed cross-lingually: it is trained on the source language, including candidate generation (Aspell, word embeddings, n-gram probabilities), then at prediction time, it is applied to the raw data in the target language. The best source language to transfer from is chosen based on empirical results on the training data.

MaChAmp* (van der Goot, 2021)

The team of CL-MoNoise also used a sequence labeling approach, in learning (character) transformations of each original word to its normalization. Scores are low on datasets that include capitalization correction, as this is not properly included in the current transformation algorithm (everything is lower-cased beforehand). The method performed much better when based on XLM-R (Conneau et al., 2020) than mBERT. Potential improvements could be gained by exploiting the multi-task capabilities of MaChAmp (van der Goot et al., 2021).

4 Intrinsic Evaluation

4.1 Intrinsic Metric

A wide variety of evaluation metrics have been used to evaluate lexical normalization performance, including accuracy over OOV words, F1 score, BLEU, word error rate, and character error rate. We choose to use Error Reduction Rate (ERR) (van der Goot, 2019b) as our main metric. ERR is the word-level accuracy normalized for the percentage of words that are in need of normalization. To calculate ERR, we use the word-level accuracy, and the percentage of words that are not normalized in the annotation:

$$ERR = \frac{\%\text{accuracy} - \%\text{words\_not\_normed}}{100 - \%\text{words\_not\_normed}}$$

We choose to use ERR instead of word-level accuracy to be able to compare (and combine) scores across datasets, since different numbers of candidates are in need of normalization. An accuracy of 93% might be a very good score on one dataset, whereas on another dataset a normalization model which scores 93% might be completely useless. The ERR will normally have a value between 0% and 100%. A negative ERR indicates that the system normalizes more words wrongly than correctly. The Leave-As-Is baseline (Section 3), which simply returns the input words, will thus by definition score an ERR of 0.0. For a more in-depth discussion of evaluation metrics for normalization and ERR, we refer the reader to Section 5.1 of van der Goot (2019b).

Team	Avg.	DA	DE	EN	ES	HR	ID-EN	IT	NL	SL	SR	TR	TR-DE
ÚFAL-2	**67.3**	68.7	**66.2**	**75.6**	59.3	**67.7**	**67.2**	**47.5**	**63.6**	**80.1**	**74.6**	**68.6**	**68.6**
ÚFAL-1	66.2	**70.3**	65.7	73.8	55.9	67.3	66.2	42.6	62.7	79.9	73.6	**68.6**	68.2
HEL-LJU-2*	53.6	56.7	59.8	62.1	35.6	56.2	55.3	35.6	45.9	67.0	66.4	51.2	51.2
HEL-LJU-1*	51.8	56.7	58.0	60.8	33.7	51.8	53.3	35.6	44.0	66.0	60.3	49.5	52.0
MoNoise	49.0	51.3	47.0	74.4	45.5	52.6	59.8	21.8	49.5	61.9	59.6	28.2	36.7
TrinkaAI-2	43.8	45.9	47.3	66.0	**61.3**	41.3	56.4	15.8	45.7	59.5	44.5	15.5	25.8
TrinkaAI-1	43.6	45.9	47.3	64.5	**61.3**	41.3	56.4	15.8	45.7	59.5	44.5	15.5	25.8
thunderml-1	43.4	46.5	46.6	64.1	60.3	40.1	59.1	11.9	44.1	59.3	44.5	15.9	29.0
team-2	40.7	48.1	46.1	63.7	21.0	40.4	59.3	13.9	43.7	60.6	46.1	15.9	29.7
learnML-2	40.3	40.5	43.7	61.6	56.6	38.1	56.2	5.9	42.8	58.3	40.0	14.4	25.7
maet-1	40.1	48.1	46.1	63.9	21.0	40.4	59.3	5.9	43.7	60.6	46.1	15.9	29.7
MFR	38.4	49.7	32.1	64.9	25.6	36.5	61.2	16.8	37.7	56.7	42.6	14.5	22.1
thunderml-2	36.5	−4.4	46.0	63.5	21.6	41.0	58.4	12.9	45.0	60.4	46.9	17.4	29.3
team-1	36.5	−4.4	46.0	63.5	21.6	41.0	58.4	12.9	45.0	60.4	46.9	17.4	29.3
CL-MoNoise*	12.1	7.3	16.6	4.1	5.0	26.4	2.4	0.0	16.2	8.8	20.1	17.6	20.2
maet-2	7.3	2.2	4.3	21.7	0.0	9.9	19.2	0.0	2.1	18.4	8.1	0.8	1.2
learnML-1	7.3	2.2	4.3	21.7	0.0	9.9	19.2	0.0	2.1	18.4	8.1	0.8	1.2
BLUE-2	6.7	49.7	−1.9	26.8	−9.4	−10.1	−7.2	−31.7	−2.1	−1.0	42.6	10.0	15.0
BLUE-1	5.2	49.7	−1.9	26.8	−10.2	−9.9	−7.2	−31.7	−2.1	−1.1	42.6	1.0	6.6
LAI	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MaChAmp*	−21.3	−88.9	−93.4	51.0	25.4	42.6	39.5	−312.9	1.5	56.8	39.4	−12.7	−3.4

Table 4: ERR on the test data (%). Negative values indicate that the system normalizes more words wrongly than correctly. Gray rows indicate baseline systems provided by the organizers. * Teams including an organizer.

The official winner of the shared task is the highest-scoring team (macro-averaged over all datasets/languages) with an open-source implementation.

4.2 Results: Intrinsic Evaluation

The main results of the shared task are shown in Table 4. All submissions except MaChAmp beat the LAI baseline, and most of them beat the MFR baseline, which turned out to be a strong baseline.

The overall winner of the shared task was ÚFAL-2, which beat the second-best team by a staggering 13 points. Recall that this method is based on ByT5, a transformer-based encoder–decoder byte-level system. Of particular note is that while recent neural approaches (Lourentzou et al., 2019; Muller et al., 2019) had not been clearly superior to the baseline MoNoise on English, ÚFAL-2 surpassed the baseline method by an appreciable margin.

The second-best team was HEL-LJU, who improved over MoNoise by more than four points. The authors report that character-level NMT provided lower results than their SMT approach, even with backtranslation.

The remaining teams mostly used either token classification or encoder–decoder approaches.

The results of this shared task showed that for state-of-the-art results one needs: (1) pre-trained models; (2) an encoder–decoder architecture over bytes or characters; and (3) synthetic task-related data.

One thing in common for the datasets with lower results in Table 4 (e.g., ES, NL, TR, and TR-DE) is that they include annotation for capitalization. Unsurprisingly, smaller datasets also tend to result in lower scores in general.

5 Extrinsic Evaluation

We perform a main extrinsic evaluation of the impact of normalization on dependency parsing using Universal Dependency annotations (Nivre et al., 2020). See Section 5.2 for test set details. We used version 2.8 of all treebanks, and syntactically-split multiword tokens and empty nodes (ellipsis) are undone with ud-conversion-tools.[6] We trained MaChAmp (van der Goot et al., 2021) with default settings and XLM-R embeddings (Conneau et al., 2020). We use the largest canonical treebank of each of the respective languages as the source domain, and attempt to improve performance on the target domain by normalizing data

[6] https://github.com/bplank/ud-conversion-tools

first. In other words, we take the input text, use a normalization system to get the normalized version, and pass this to the parser as input. The parsers are not trained on social media data, to evaluate the impact of normalization on parsing in a domain-shift scenario. Because the normalized version should be closer to the canonical training data of the parser, performance is expected to improve compared to using the input directly. The training treebanks are UD_English-EWT (Silveira et al., 2014), UD_German-GSD (Brants et al., 2004), UD_Italian-ISDT (Bosco et al., 2014), and UD_Turkish-IMST (Sulubacak et al., 2016). For the extrinsic evaluation, we report the average over three runs with different random seeds (only the parser is retrained, not the normalization models).

5.1 Alignment-aware Parsing Metrics

Because our definition of the lexical normalization task includes the splitting and merging of tokens (namely "one-to-many" and "many-to-one" replacements, cf. Section 1), the standard evaluation for dependency parsing has to be adapted for our purposes. Specifically, we compute the token-level labeled attachment score (LAS) and unlabeled attachment score (UAS) after aligning predicted tokens to gold tokens. We refer to these alignment-aware metric variants as *aligned* LAS and UAS (i.e., *a*-LAS and *a*-UAS), respectively. Because "many-to-one" replacements are relatively rare, we cannot check when they are correct (normalization annotation is not available for most treebanks) and they are non-trivial to include in the aligned evaluation, and hence we decided to undo these in the system outputs, and use the original input instead. For the "one-to-many" replacements, we check whether one of the words in the split is connected correctly. All incorrect words in the 'many' are thus simply ignored. It should be noted that this can give an advantage to systems that split, and we thus suggest that this metric is always reported with the number of splits. Furthermore, we assume none of the teams made use of this shortcoming in the metric, as they were unaware of these details.

5.2 Test Sets and Metric

We employ *a*-LAS as the main metric for extrinsic evaluation, and also report *a*-UAS for the sake of completeness. Each system was tested on 7 dependency parsing treebanks consisting of posts from Twitter in 4 different languages:

- German (DE): TweeDe (Rehbein et al., 2019);
- English (EN): AAE (Blodgett et al., 2018), MoNoise (van der Goot and van Noord, 2018), and Tweebank2 (Liu et al., 2018);
- Italian (IT): PoSTWITA (Sanguinetti et al., 2018) and TWITTIRO (Cignarella et al., 2019);
- Turkish (TR): IWT151 (Pamay et al., 2015; Sulubacak and Eryiğit, 2018).

5.3 Results: Impact on Parsing

The main results (*a*-LAS) are reported in Table 5. Although most submissions outperform the LAI baseline, it becomes clear that lexical normalization is only a step towards closing the gap in performance on canonical data, as performance is still far from the average LAS on our training sets (79 LAS). This is confirmed by the scores of using the manually-annotated (gold) normalization.[7] The best performing model scores 1.72 LAS points higher than the LAI baseline. Compared to normalization performance (Table 4), the baselines (LAI and MFR) rank highly, especially for tr-iwt151 and en-tweebank2, which is probably because they have less risk of over-normalization, and some of the treebanks might need only very little normalization (there is also an abundance of canonical data to be found on Twitter). The largest gains compared to the LAI baseline are obtained on the en-monoise treebank, probably because this treebank contains data filtered to contain data in need of normalization. In the gold-standard annotations, 31.33% of the words are normalized, compared to 1.04% for en-tweebank2.

The full *a*-UAS scores can be found in the appendix (Table 6). In general, performance is approximately 7–10 points higher than LAS (absolute), and differences between teams are smaller. Interestingly, the ranking is slightly different there, with HEL-LJU and MFR ranking higher, and CL-MoNoise ranking lower. Overall, the results confirm that the best normalization systems (by ÚFAL and HEL-LJU) also result in the highest observed parsing improvements on these social media treebank test sets. Once again, these two teams outperform the previous state-of-the-art system (i.e., MoNoise).

[7] Note that MoNoise and Tweebank2 are the only treebanks where a normalization layer is fully annotated, so we report Gold results for these treebanks only.

Treebank	Avg.		de-tweede		en-aae		en-monoise		en-tweebank2		it-postwita		it-twittiro		tr-iwt151	
ÚFAL-2	**64.2**	37	**73.6**	5	**62.7**	53	**58.6**	33	59.1	66	**68.3**	13	72.2	5	54.7	90
ÚFAL-1	64.0	37	**73.6**	3	62.2	50	57.9	33	59.0	65	**68.3**	14	72.2	5	54.8	95
HEL-LJU-2*	63.7	10	73.5	3	60.6	20	56.3	4	60.3	15	68.1	3	72.3	0	55.0	28
HEL-LJU-1*	63.7	14	73.5	3	60.5	18	56.3	9	60.3	18	68.2	3	**72.5**	0	54.9	48
MoNoise	63.4	21	73.2	5	62.3	40	56.8	18	58.9	46	67.6	3	70.7	0	54.6	35
MFR	63.3	16	72.9	5	60.3	32	56.7	15	**60.3**	37	67.3	3	70.7	0	54.9	25
TrinkaAI-2	63.1	33	72.9	7	60.2	40	56.6	19	59.9	39	67.0	7	71.1	0	54.2	119
TrinkaAI-1	63.1	33	72.9	7	60.2	40	56.6	19	59.9	39	67.0	7	71.1	0	54.2	119
maet-1	63.1	27	72.8	3	59.4	40	56.6	24	59.8	44	67.4	10	71.1	0	54.5	74
team-2	63.0	27	72.8	3	59.4	40	56.6	24	59.8	44	67.2	4	70.9	3	54.5	74
thunderml-2	63.0	33	72.7	3	59.6	42	56.7	28	59.3	44	67.3	4	71.4	1	54.2	112
team-1	63.0	33	72.7	3	59.6	42	56.7	28	59.2	44	67.3	4	71.4	1	54.2	112
thunderml-1	63.0	34	72.5	6	59.3	49	56.7	24	59.9	46	67.1	5	71.0	0	54.1	112
learnML-2	62.9	30	72.3	8	59.0	44	56.2	31	60.0	45	67.0	6	71.2	0	54.5	79
CL-MoNoise*	62.7	24	72.7	0	60.9	0	55.3	0	58.5	0	66.5	99	70.1	50	55.0	20
BLUE-2	62.5	0	72.6	0	59.6	0	54.2	0	59.8	0	66.7	0	70.0	0	54.8	0
BLUE-1	62.5	0	72.6	0	59.6	0	54.2	0	59.8	0	66.7	0	70.0	0	54.8	0
LAI	62.5	0	72.7	0	59.2	0	53.7	0	60.0	0	66.5	0	70.1	0	**55.0**	0
maet-2	62.2	0	72.7	0	58.5	0	52.9	0	60.0	0	66.5	0	70.0	0	55.0	0
learnML-1	62.2	0	72.7	0	58.5	0	52.9	0	60.0	0	66.5	0	70.0	0	55.0	0
MaChAmp*	61.9	43	71.3	7	60.8	37	54.6	43	58.0	52	64.7	2	69.8	0	54.1	162
Gold	—		—		—		60.8		60.4		—		—		—	

Table 5: a-LAS scores (%) and the number of splits (in smaller font) for each dataset. Gray rows indicate baseline systems provided by the organizers. * Teams including an organizer. The final row ("Gold") indicates performance with gold-standard normalization.

5.4 Results: Impact on POS tagging

Additionally, we calculated the POS accuracies using the same heuristic as described in Section 5.1 (i.e., a-POS), and present full results in the appendix (Table 7). Again, we see some changes in the ranking of the teams, and performance improvements are slightly more moderate compared to a-LAS. The baselines score highest on en-tweebank2 (MFR) and tr-iwt151 (LAI), and the highest gains are again obtained on the en-monoise treebank.

6 Conclusions

With MULTILEXNORM, we have developed a multilingual benchmark for lexical normalization consisting of previously-created datasets spanning 12 language variants. We proposed a standard evaluation metric, and both intrinsic and extrinsic evaluation via dependency parsing and POS tagging. We hosted a shared task with this new benchmark, which enabled comparison of performance of 21 models (18 submissions by participants, and 3 in-house baselines). The results of the shared task show that the previous state of the art on lexical normalization is outperformed by a large margin. The extrinsic evaluation on dependency parsing and POS tagging shows that lexical normalization is beneficial (with improvements in a-LAS and a-UAS of up to +1.72 and +0.85, and improvements in a-POS of up to +1.54, respectively), but there is still a performance gap compared to the performance levels observed on canonical data. We hope that the proposed benchmark will lead to more research in multilingual normalization, and more transparent and fairer comparisons. All submissions, evaluation scripts, and baseline models are available in the shared task repository.

Acknowledgements

B.M. was funded by the French Research Agency via the ANR ParSiTi project (ANR-16-CE33-0021).

References

Inaki Alegria, Nora Aranberri, Víctor Fresno, Pablo Gamallo, Lluis Padró, Inaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. 2013. Introducción a la tarea compartida Tweet-Norm 2013: Normalización léxica de tuits en Español. In *Tweet-Norm@SEPLN*, pages 1–9.

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for SMS text normalization. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 33–40, Sydney, Australia. Association for Computational Linguistics.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan. Asian Federation of Natural Language Processing.

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China. Association for Computational Linguistics.

Tyler Baldwin and Yunyao Li. 2015. An in-depth analysis of the effect of text normalization in social media. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 420–429, Denver, Colorado. Association for Computational Linguistics.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the evalita 2016 sentiment polarity classification task. In *Proceedings of Evalita 2016*.

Anab Maulana Barik, Rahmad Mahendra, and Mirna Adriani. 2019. Normalization of Indonesian-English code-mixed Twitter data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 417–424, Hong Kong, China. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Irshad Bhat, Riyaz A. Bhat, Manish Shrivastava, and Dipti Sharma. 2018. Universal Dependency parsing for Hindi-English code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 987–998, New Orleans, Louisiana. Association for Computational Linguistics.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency parsing for African-American and mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

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.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The EVALITA 2014 dependency parsing task. In *EVALITA 2014 Evaluation of NLP and Speech Tools for Italian*, pages 1–8. Pisa University Press.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2):55–63.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a german corpus. *Research on language and computation*, 2(4):597–620.

Ana-Maria Bucur, Adrian Cosma, and Liviu P. Dinu. 2021. Sequence-to-sequence lexical normalization with multilingual transformers. In *Proceedings of the 7th Workshop on Noisy User-generated Text (W-NUT 2021)*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Özlem Çetinoğlu. 2016. A Turkish-German code-switching corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4215–4220, Portorož, Slovenia. European Language Resources Association (ELRA).

Alessandra Teresa Cignarella, Cristina Bosco, and Paolo Rosso. 2019. Presenting TWITTIRÒ-UD: An Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 190–197, Paris, France. Association for Computational Linguistics.

Talha Çolakoğlu, Umut Sulubacak, and Ahmet Cüneyd Tantuğ. 2019. Normalizing non-canonical Turkish texts using machine translation approaches. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 267–272, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

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.

Anne Dirkson, Suzan Verberne, and Wessel Kraaij. 2019. Lexical normalization of user-generated medical text. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 11–20, Florence, Italy. Association for Computational Linguistics.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Tomaž Erjavec, Darja Fišer, Jaka Čibej, Špela Arhar Holdt, Nikola Ljubešić, and Katja Zupan. 2017. CMC training corpus Janes-Tag 2.0. Slovenian language resource repository CLARIN.SI.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017a. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017b. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Divesh Kubal and Apurva Nagvenkar. 2021. Multilingual sequence labeling approach to solve lexical normalization. In *Proceedings of the 7th Workshop on Noisy User-generated Text (W-NUT 2021)*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chen Li and Yang Liu. 2015. Joint pos tagging and text normalization for informal text. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing tweets into Universal Dependencies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 965–975, New Orleans, Louisiana. Association for Computational Linguistics.

Nikola Ljubešić, Tomaž Erjavec, Maja Miličević, and Tanja Samardžić. 2017a. Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Tomaž Erjavec, Maja Miličević, and Tanja Samardžić. 2017b. Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2014. TweetCaT: a tool for building Twitter corpora of smaller languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2279–2283, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak, and Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 371–378, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Nikola Ljubešić and Denis Kranjčić. 2015. Discriminating between closely related languages on twitter. *Informatica*, 39(1):1.

Ismini Lourentzou, Kabir Manghnani, and ChengXiang Zhai. 2019. Adapting sequence to sequence models for text normalization in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 335–345.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

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.

Dong Nguyen, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in NLP: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Kemal Oflazer. 1994. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2):137–148.

NHJ Oostdijk, Martin Reynaert, Veronique Hoste, H Heuvel, O De Clercq, and EP Sanders. 2014. Sonar nieuw media corpus.

Tuğba Pamay, Umut Sulubacak, Dilara Torunoğlu-Selamet, and Gülşen Eryiğit. 2015. The annotation process of the ITU web treebank. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 95–101, Denver, Colorado, USA. Association for Computational Linguistics.

Deana L. Pennell and Yang Liu. 2014. Normalization of informal text. *Computer Speech & Language*, 28(1):256–277.

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.

Ines Rehbein, Josef Ruppenhofer, and Bich-Ngoc Do. 2019. tweeDe – a Universal Dependencies treebank for German tweets. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 100–108, Paris, France. 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 7th Workshop on Noisy User-generated Text (W-NUT 2021)*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli, and Fabio Tamburini. 2018. PoSTWITA-UD: an Italian Twitter treebank in Universal Dependencies. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tatjana Scheffler. 2014. A German Twitter snapshot. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2284–2289, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yves Scherrer and Nikola Ljubešić. 2021. Sesame Street to Mount Sinai: BERT-constrained character-level Moses models for multilingual lexical normalization. In *Proceedings of the 7th Workshop on Noisy User-generated Text (W-NUT 2021)*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Helmut Schmid. 1994. Part-of-speech tagging with neural networks. In *15th International Conference on Computational Linguistics, COLING 1994, Kyoto, Japan, August 5-9, 1994*, pages 172–176.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Sarah Schulz, Guy De Pauw, Orphée De Clercq, Bart Desmet, Veronique Hoste, Walter Daelemans, and Lieve Macken. 2016. Multimodular text normalization of dutch user-generated content. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(4):1–22.

Youri Schuur. 2020. Normalization for Dutch for improved pos tagging. Master's thesis, University of Groningen.

Uladzimir Sidarenka. 2019. *Sentiment analysis of German Twitter*. Ph.D. thesis, Universität Potsdam.

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*.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Richard Sproat, Alan W Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer speech & language*, 15(3):287–333.

Umut Sulubacak and Gülşen Eryiğit. 2018. Implementing universal dependency, morphology, and multiword expression annotation standards for turkish language processing. *Turkish Journal of Electrical Engineering & Computer Sciences*, 26(3):1662–1672.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Rob van der Goot. 2019a. MoNoise: A multi-lingual and easy-to-use lexical normalization tool. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 201–206, Florence, Italy. Association for Computational Linguistics.

Rob van der Goot. 2019b. *Normalization and Parsing Algorithms for Uncertain Input.* Ph.D. thesis, University of Groningen.

Rob van der Goot. 2021. CL-MoNoise: Cross-lingual lexical normalization. In *Proceedings of the 7th Workshop on Noisy User-generated Text (W-NUT 2021)*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rob van der Goot and Özlem Çetinoğlu. 2021. Lexical normalization for code-switched data and its effect on POS tagging. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2352–2365, Online. Association for Computational Linguistics.

Rob van der Goot, Barbara Plank, and Malvina Nissim. 2017. To normalize, or not to normalize: The impact of normalization on part-of-speech tagging. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 31–39, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Rob van der Goot and Gertjan van Noord. 2017. Parser adaptation for social media by integrating normalization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 491–497, Vancouver, Canada. Association for Computational Linguistics.

Rob van der Goot and Gertjan van Noord. 2018. Modeling input uncertainty in neural network dependency parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4984–4991, Brussels, Belgium. Association for Computational Linguistics.

Rob van der Goot, Rik van Noord, and Gertjan van Noord. 2018. A taxonomy for in-depth evaluation of normalization for user generated content. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Cynthia Van Hee, Marjan Van de Kauter, Orphée De Clercq, Els Lefever, Bart Desmet, and Veronique Hoste. 2017. Noise or music? investigating the usefulness of normalisation for robust sentiment analysis on social media data. *Traitement Automatique Des Langues*, 58(1):63–87.

Ke Xu, Yunqing Xia, and Chin-Hui Lee. 2015. Tweet normalization with syllables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 920–928, Beijing, China. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *arXiv preprint arXiv:2105.13626*.

Yi Yang and Jacob Eisenstein. 2013. A log-linear model for unsupervised text normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 61–72, Seattle, Washington, USA. Association for Computational Linguistics.

Katja Zupan, Nikola Ljubešić, and Tomaž Erjavec. 2019. How to tag non-standard language: Normalisation versus domain adaptation for slovene historical and user-generated texts. *Natural Language Engineering*, 25(5):651–674.

A MULTILEXNORM Data Statement

Following Bender and Friedman (2018), we present statements for MULTILEXNORM data.

A. CURATION RATIONALE

- **Danish**: We collected data from Twitter by querying the Twitter API using the following emotion-related keywords: *frygt*, *glæd*, *kærlighed*, *overraskelse*, *racistisk*, *sjov*, *smerte*, *tristhed*. Tweets were collected in 2019–2020. We also scraped all Arto pages from the Internet Wayback machine (archive.org), and extracted the user-generated content from the HTML with a script. We then applied filtering on the combination of this data, leaving only sentences which were classified as Danish with a confidence of at least 0.885 by the FastText language classifier (Joulin et al., 2017b) and contain at least 3 words in the Danish Aspell dictionary (which are not in the English dictionary), and contain at least 2 words not in the Danish Aspell dictionary.

- **German**: To create this corpus, we randomly sampled 10,000 messages from the German Twitter Snapshot (GTS; Scheffler, 2014)—a collection of 24 million tweets, which were gathered in April 2013 by permanently tracking a list of 397 frequent German words via the Twitter Streaming API and subsequently filtered with `langid.py` (Lui and Baldwin, 2012). We analyzed all tokens of the sample with `TreeTagger` (Schmid, 1994) and `hunspell`. Afterwards, two human experts annotated all words that any of these tools considered as out-of-vocabulary (OOV) and that appeared at least twice in the selected microblogs or belonged to a set of 1,000 randomly-chosen hapax legomena. Finally, we only left tweets that contained words annotated as *spelling deviation* by either of the experts, resulting in a total of 1,492 messages.

- **English**: Tweets were collected using the Twitter Streaming API over the period 23–29 May, 2014, and then filtered by `langid.py` (Lui and Baldwin, 2012) to remove non-English tweets. To ensure that tweets had a high likelihood of requiring lexical normalization, tweets with less than 2 non-standard words (i.e. words not occurring in the SCOWL dictionary) were filtered out.

- **Spanish**: To maximize the chances of getting tweets in the Spanish language, tweets were collected through Twitter's streaming API by restricting the search to a geographical bounding box within Spain but excluding bilingual regions. The selected geographic area forms a rectangle with Guadalajara (coordinates: 41, -2) as the northeasternmost point and Cadiz (coordinates: 36.5, -6) as the southwesternmost point. The resulting collection with over 227K tweets was filtered to keep only tweets identified by Twitter as having been written in Spanish (i.e. 'lang' field set to 'es'), and further sampling was done to make it manageable for manual labeling.

- **Croatian**: The dataset is a subset of the large Croatian Twitter crawl harvested with TweetCat (Ljubešić et al., 2014) between 2013 and 2016. It contains a similar amount of standard and non-standard data, and non-standard data was oversampled from the original data collection. The standardness level of the data was predicted via feature-based machine learning (Ljubešić et al., 2015). Discrimination between Croatian and Serbian tweets was performed with a dedicated supervised classifier (Ljubešić and Kranjčić, 2015).

- **Indonesian-English**: Barik et al. (2019) collected Indonesian-English code-mixed tweets using the Twitter search API. First, they compiled a list of Indonesian and English stopwords (100 for each language), based on frequent word lists from Wiktionary.[8] The stopwords were then used as search queries. In order to obtain code-mixed tweets, the "language" parameter in the search query was set to be constrastive to the language of the stopword used. For example, the "language" parameter is set to English when an Indonesian stopword is used as a query, and vice versa. To minimize chance that tweets contain any word from local indigenous or other languages, the "location" parameter in the search query is restricted to only Jakarta and Bandung (the two largest cities in Indonesia). In total, 49,647 tweets were collected. Two human annotators

[8] https://en.wiktionary.org/wiki/Wiktionary:Frequencylists

labeled a sample of 825 tweets from the larger collection. The annotators were instructed to tokenize tweets into a list of word segments, and then provide the lowercase normalized form for each segment. A segment can be a single or multi-token word, untokenized proper name, hyperlink, emoticon, or Twitter special term (i.e., hashtag or mention).

- **Italian**: The dataset is a subset of the data from Sanguinetti et al. (2018) (v. 2.1), which in turn is a subset of SENTIPOLC (Barbieri et al., 2016) and SentiTUT (Bosco et al., 2013). Tweets were mostly collected during the period 2011–2012, and have been filtered based on keywords about politics, in addition to a small subset from the random Twitter API stream. Tweets that contain ≥ 3 out-of-vocabulary words (i.e., not in the Aspell dictionary for Italian, or either a URL, hashtag, username, or text consisting of punctuation-only) were filtered out to ensure a basic density of non-standard language for further annotation. Moreover, a small list of proper nouns was added to the vocabulary, taken from the most frequent out-of-vocabulary words in the dataset.

- **Dutch**: We took the data from the SoNaR Nieuwe Media Corpus (Oostdijk et al., 2014) as a starting point, and selected sentences which contain at least 3 words which are not in the Aspell dictionary for Dutch. We originally took 500 sentences from each sub-domain (SMS, chats, and tweets), and then removed all sentences which were completely written in another language (i.e., Frisian, Afrikaans, English, or Spanish).

- **Slovenian**: The dataset is a subset of a large Slovenian Twitter crawl harvested with TweetCat (Ljubešić et al., 2014) between 2013 and 2016. It contains a similar amount of standard and non-standard data, and non-standard data was oversampled from the original data collection. The standardness level of the data was predicted via feature-based machine learning (Ljubešić et al., 2015).

- **Serbian**: The dataset is a subset of a large Serbian Twitter crawl harvested with TweetCat (Ljubešić et al., 2014) between 2013 and 2016. It contains a similar amount of standard and non-standard data, and non-standard data was oversampled from the original data collection. The standardness level of the data was predicted via feature-based machine learning (Ljubešić et al., 2015). Discrimination between Croatian and Serbian tweets was performed with a dedicated supervised classifier (Ljubešić and Kranjčić, 2015).

- **Turkish-German**: The code-switched dataset is derived by filtering tweets labeled as Turkish and German according to Twitter's language ID assignment. Turkish tweets were collected in 2015 and German tweets during 2009–2011. To identify mixed German–Turkish tweets, we mainly used morphological analyzers (Oflazer, 1994; Schmid et al., 2004) as filters. Manual filtering followed the automatic filtering, resulting in the final dataset. The raw tweets were manually tokenized, normalized and segmented (Çetinoğlu, 2016). In addition, usernames and URLs were anonymized as @username and [url], respectively, and language IDs for each token were added. Adapting the dataset to the normalization task was performed in van der Goot and Çetinoğlu (2021).

B. LANGUAGE VARIETY All of the datasets consist of social media variants of the standard languages, and are not bound by a regional standard (i.e., no distinction is made between `en_us` or `en_gb`).

C. SPEAKER DEMOGRAPHIC The speaker demographics are unknown. For some of the collected data this might have been available, but it is not shared on purpose (for privacy reasons).

D. ANNOTATOR DEMOGRAPHIC

- **Danish**: Two native speakers of Danish. Both were higher-education students (male and female), between the age of 20 and 30.

- **German**: An undergraduate (native German speaker studying computational linguistics), and a PhD student (Belarusian Germanist pursuing a degree in computational linguistics).

- **English**: 12 interns and employees at IBM Research Australia were involved in the data annotation. All annotators had a high level of

Treebank	Avg.		de-tweede		en-aae		en-monoise		en-tweebank2		it-postwita		it-twittiro		tr-iwt151	
ÚFAL-2	**74.0**	37	83.2	5	**72.5**	53	**68.4**	33	69.3	66	78.0	13	81.0	5	65.5	90
HEL-LJU-2*	74.0	10	**83.3**	3	71.2	20	67.2	4	**71.1**	15	77.9	3	81.3	0	65.7	28
HEL-LJU-1*	73.9	14	**83.3**	3	70.9	18	67.4	9	71.1	18	78.0	3	**81.3**	0	65.7	48
ÚFAL-2	73.8	37	83.2	3	71.9	50	68.0	33	69.3	65	**78.0**	14	80.8	5	65.5	95
MFR	73.6	16	82.9	5	70.7	32	67.4	15	71.0	37	77.2	3	80.1	0	65.7	25
MoNoise	73.4	21	83.1	5	71.9	40	67.0	18	68.6	46	77.3	3	80.0	0	65.5	35
TrinkaAI-2	73.3	33	82.7	7	70.1	40	67.4	19	70.7	39	77.0	7	80.5	0	65.0	119
TrinkaAI-1	73.3	33	82.7	7	70.1	40	67.4	19	70.7	39	77.0	7	80.5	0	65.0	119
thunderml-2	73.3	33	82.3	3	69.9	42	67.3	28	70.0	44	77.5	4	80.9	1	65.2	112
team-1	73.3	33	82.3	3	69.9	42	67.3	28	70.0	44	77.5	4	80.9	1	65.2	112
maet-1	73.2	27	82.6	3	69.3	40	67.2	24	70.5	44	77.4	10	80.3	0	65.4	74
team-2	73.2	27	82.6	3	69.3	40	67.2	24	70.5	44	77.2	4	80.3	3	65.4	74
thunderml-1	73.1	34	82.4	6	68.9	49	67.3	24	70.6	46	77.0	5	80.7	0	65.0	112
LAI	73.1	0	82.6	0	70.6	0	65.3	0	70.9	0	76.7	0	79.8	0	65.8	0
learnML-2	73.0	30	82.2	8	68.7	44	66.5	31	70.7	45	77.0	6	80.6	0	65.3	79
BLUE-2	73.0	0	82.4	0	70.5	0	65.6	0	70.4	0	76.9	0	79.6	0	65.6	0
BLUE-1	73.0	0	82.4	0	70.5	0	65.6	0	70.4	0	76.9	0	79.6	0	65.6	0
maet-2	72.9	0	82.6	0	69.7	0	65.1	0	70.9	0	76.7	0	79.7	0	65.8	0
learnML-1	72.9	0	82.6	0	69.7	0	65.1	0	70.9	0	76.7	0	79.7	0	65.8	0
CL-MoNoise*	72.8	24	82.5	0	71.8	0	65.5	0	67.9	0	76.6	99	79.8	50	**65.8**	20
MaChAmp*	72.6	43	81.8	7	71.0	37	65.5	43	68.4	52	76.3	2	80.3	0	65.0	162
Gold	—		—		—		69.1		71.2		—		—		—	

Table 6: *a*-UAS scores (%) and the number of splits (in smaller font) for each dataset. Gray rows indicate baseline systems provided by the organizers. * Teams including an organizer. The final row ("Gold") indicates performance with gold-standard normalization.

English proficiency (IELTS 6.0+) and were reasonably familiar with Twitter data.

- **Spanish**: Nine native speakers of Spanish. Eight male and one female with ages ranging from 30 to 60. All annotators had a background in natural language processing and were familiar with the Twitter platform.
- **Croatian**: Three native speakers of Croatian, all linguists with an MA degree, trained in data annotation.
- **Indonesian-English**: Two native speakers of Indonesian, fluent in English. Both annotators were 22 years old at the time of the annotation.
- **Italian**: Four native speakers of Italian, all male, between the age of 20 and 38, from a variety of Italian regions (i.e., Veneto, Tuscany, Liguria, and Apulia). All annotators had a background in natural language processing and were familiar with the Twitter platform.
- **Dutch**: The main annotator was a native Dutch Information Science master student (male, age range 20–25). The second annotator (for agreement scores) was a native Dutch male PhD student in NLP, age 27.
- **Slovenian**: Five native speakers of Slovenian, all master-level students of language-related studies.
- **Serbian**: Two native speakers of Serbian, all linguists with an MA degree, trained in data annotation.
- **Turkish-German**: The annotators were three Turkish-German bilinguals born and raised in Germany. They have studied computational linguistics and their age ranged from 20 to 25.

E. SPEECH SITUATION The data is not spoken. However, input methods might have changed over time. A tweet collected from 2012 was less likely to be produced with a spell checker compared to one collected from 2020.

F. TEXT CHARACTERISTICS The genre is not bound topic- or content-wise. However, all inputs are shorter than 280 characters, and most of

Treebank	Avg.	de-tweede	en-monoise	en-tweebank2	it-postwita	it-twittiro	tr-iwt151
ÚFAL-2	**85.5**	**87.6**	**84.1**	80.5	**86.6**	88.7	85.5
ÚFAL-1	85.4	87.5	83.7	80.4	**86.6**	**88.8**	85.5
HEL-LJU-1*	85.1	87.4	81.4	81.6	86.3	88.6	85.6
HEL-LJU-2*	85.1	87.4	81.3	81.6	86.2	88.5	85.6
MoNoise	84.9	87.3	82.2	81.6	85.4	87.3	85.5
thunderml-2	84.9	87.1	82.7	81.2	85.7	87.6	85.1
team-1	84.9	87.1	82.7	81.2	85.7	87.6	85.1
team-2	84.8	87.1	82.2	81.3	85.5	87.7	85.3
MFR	84.8	87.1	82.1	**81.7**	85.3	87.2	85.6
maet-1	84.8	87.1	82.2	81.3	85.5	87.5	85.3
learnML-2	84.8	87.1	82.2	81.5	85.2	87.5	85.3
TrinkaAI-2	84.7	87.2	82.0	81.4	85.4	87.3	85.1
TrinkaAI-1	84.7	87.2	82.0	81.4	85.4	87.3	85.1
thunderml-1	84.7	86.9	82.2	81.4	85.5	87.5	85.0
CL-MoNoise*	84.2	86.9	80.1	81.5	84.4	86.6	85.6
BLUE-2	84.2	87.0	79.8	81.5	84.4	86.8	85.6
BLUE-1	84.2	87.0	79.8	81.5	84.4	86.8	85.6
LAI	84.0	87.1	78.8	81.4	84.3	86.5	**85.7**
maet-2	83.9	87.1	78.7	81.4	84.3	86.5	**85.7**
learnML-1	83.9	87.1	78.7	81.4	84.3	86.5	**85.7**
MaChAmp*	82.0	84.6	80.2	80.8	80.0	83.0	83.8
Gold	—	—	85.5	81.6	—	—	—

Table 7: *a*-POS accuracy (%) for each dataset. Gray rows indicate baseline systems provided by the organizers. * Teams including an organizer. The final row ("Gold") indicates performance with gold-standard normalization.

them shorter than 140 characters (Twitter increased the maximum tweet length in September 2017).

I. PROVENANCE APPENDIX The data is released under a CC-BY-SA license.

B *a*-UAS Scores

We report *a*-UAS scores in Table 6.

C *a*-POS Accuracies

We report *a*-POS accuracy values in Table 7. Note that the en-aae treebank is not included here because it has no POS annotation.