Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets

Jaka Čibej[1], Darja Fišer[1,2], Tomaž Erjavec[2]
[1] Dept. of Translation, Faculty of Arts, University of Ljubljana
Aškerčeva 2, 1000 Ljubljana
[2] Dept. of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
E-mail: jaka.cibej@ff.uni-lj.si, darja.fiser@ff.uni-lj.si, tomaz.erjavec@ijs.si

Abstract

Online user-generated content such as posts on social media, blogs, and forums, is becoming an increasingly important source of information, as shown by numerous rapidly growing NLP fields such as sentiment analysis and data mining. However, user-generated content is well-known to contain a significant degree of noise, e.g. abbreviations, missing spaces, as well as non-standard spelling, lexis, and use of punctuation. All this hinders the effectiveness of NLP tools when processing such data, and to overcome this obstacle, data normalisation is required. In this paper, we present a training set that will be used to improve the tokenisation, normalisation, and sentence segmentation of Slovene tweets. We describe some of the most Twitter-specific aspects of our annotation guidelines as well as the workflow of our annotation campaign, the goal of which was to create a manually annotated gold-standard dataset of 4,000 tweets extracted from the JANES corpus of Internet Slovene.

Keywords: normalisation, tokenisation, sentence segmentation, tweets, user-generated content

1. Introduction

With the rapid global expansion of the Internet, online user-generated content such as blogs, forums, and social media, is becoming an increasingly important source of information. The analysis of social media has become a popular research topic in a number of branches of NLP, including data mining, sentiment analysis, named entity recognition, and machine translation. However, user-generated content is well-known to contain a significant degree of noise, e.g. non-standard spelling and colloquialisms, frequent abbreviations, missing spaces and diacritics (Crystal, 2011; Eisenstein, 2013; Baldwin et al., 2013). In this regard, Slovene computer-mediated communication is no exception (Erjavec & Fišer, 2013; Zwitter Vitez & Fišer, 2015).

NLP tools trained on standard language data are less effective on noisy texts, which can be remedied through two different approaches: either by training new NLP tools on noisy data and adapting them to a particular variety of noisy language variety (see e.g. Yang & Eisenstein, 2013), or by improving the performance of existing NLP tools through data normalisation (Sproat, 2001). In the case of Slovene, a language with approximately 2 million speakers, developing new tools for its many regional and social language variants is unrealistic and unfeasible in terms of the available resources, so the logical step is to tackle noisy social media content via data normalisation.

In this paper, we present the compilation of a dataset that will be used to improve the tokenisation, normalisation and sentence segmentation of Slovene tweets in the context of the annotation of the JANES corpus of Internet Slovene (Fišer et al., 2015), a 160-million-token corpus of Slovene user-generated content containing tweets, forum posts, news site comments, and blogs.

The paper is structured as follows: in Section 2, we provide a brief overview of related work. In Section 3, we present the structure of the dataset to be annotated and the criteria used to compile it. We describe the annotation platform and the project workflow in Section 4 and then continue by describing the highlights of our annotation guidelines for sentence segmentation, tokenisation, and normalisation in Section 5. Finally, we conclude with a discussion of the results and suggestions for future work.

2. Related Work

Normalisation of Twitter content is not an uncommon task in the field of NLP. Approaches to the problem range from automatic construction of normalisation dictionaries to facilitate lexical normalisation through simple string substitution (Han et al., 2012); rule-based normalisation tackling omissions and repetitions in out-of-vocabulary tokens (Sidarenka et al., 2013; Clark & Araki, 2011); or normalisation using finite-state transducers (Porta & Sancho, 2013). In addition to normalisation models, language resources such as annotated datasets and corpora are also produced to help develop and test new normalisation systems (Alegria et al., 2014).

For Slovene, the most notable work so far for tweet normalisation is the normalisation model developed by Ljubešić et al. (2014), which aimed to improve the performance of existing Slovene text processing tools by training a character-level statistical machine translation system on a small manually validated lexicon containing pairs of original and normalised forms for the 1,000 most salient out-of-vocabulary (OOV) tokens with respect to a reference corpus. The model performed well, achieving a 69% accuracy when normalising OOV tokens, but there is still significant room for improvement. A major disadvantage of the system is that it is lexicon-based and does not take context into account when proposing normalisation. For this, an annotated corpus is required, the production of which is presented in this paper.

3. Dataset

A dataset of Slovene tweets to be manually annotated was prepared by extracting 4,000 tweets from the JANES corpus. The tweets were sampled according to their technical (T1–T3) and linguistic (L1–L3) standardness levels (Ljubešić et al., 2015), where 1 signifies a high degree of standardness and 3 a significant degree of non-standardness. For instance, a T1L3 tweet is standard from a technical perspective (punctuation, capitalisation, and use of spaces), but non-standard in linguistic terms (e.g. lexis and spelling), while a T3L1 tweet contains standard language written with e.g. no capital letters and no punctuation. A T3L3 tweet is non-standard in both regards.

- **T=1 / L=1**

Original: Bi kdo stanovanje v Kranju (Sejmišče) za 230€ na mesec? Starejše, enosobno (35m2), udobna kopalnica, visok strop, zastekljen balkon...
Standard: Bi kdo stanovanje v Kranju (Sejmišče) za 230 € na mesec? Starejše, enosobno (35 m2), udobna kopalnica, visok strop, zastekljen balkon ...
Characteristics
T: correct use of sentence-initial capitalisation and sentence-final punctuation, few missing spaces
L: completely standard lexis and spelling

- **T=3 / L=1**

Original: na sreco se motis,alkohol je v slo 100x vecji problem, primerjaj smrtnost in druzbeno skodo zaradi dovoljenih in nedovoljenih
Standard: Na srečo se motiš, alkohol je v Slo. 100x večji problem, primerjaj smrtnost in družbeno škodo zaradi dovoljenih in nedovoljenih.
Characteristics
T: no diacritics, no sentence-initial capitalisation, no sentence-final punctuation, missing spaces after punctuation
L: standard lexis and spelling

- **T=1 / L=3**

Original: Ja sej je blo to prav na koncu. Se mi je ena druga prijazna javla pa je rekla da sm prav poklical. Prvič ni šlo.
Standard: Ja saj je bilo to prav na koncu. Se mi je ena druga prijazna javila pa je rekla da sem prav poklical. Prvič ni šlo.
Characteristics
T: no missing spaces, correct use of punctuation and capitalisation
L: non-standard spelling (*sej* vs. *saj*, *blo* vs. *bilo*, *javla* vs. *javila*, *sm* vs. *sem*)

- **T=3 / L=3**

Original: jp,sis je se najbolj ziher... js sem se zarad tega 1x zastonj v portoroz peljala. mal na plazo pa tko.kaj pa 400 km :)
Standard: Jp, sis je še najbolj ziher ... Jaz sem se zaradi tega 1x zastonj v Portorož peljala. Malo na plažo pa tako. Kaj pa 400 km :)
Characteristics
T: no capitalisation (*portoroz* vs. *Portorož*), missing spaces before punctuation (*ziher... – jp,sis – tko.kaj*)
L: non-standard lexis (*ziher*, *jp*), non-standard spelling (*js* vs. *jaz*, *mal* vs. *malo*, *tko* vs. *tako*)

Figure 1. Examples of tweets with different standardness scores.

The dataset consists of four tweet categories, each contributing 1,000 tweets. The first three categories (T1L3, T3L1, and T3L3) contain tweets with the highest degree of non-standardness (either technical, linguistic, or both), while the last (T1L1) contains tweets that show next to no signs of non-standardness. Examples for each of these categories are shown in Figure 1.
The sampled tweets were automatically tokenised, segmented into sentences (Erjavec et al., 2005) and normalised (Ljubešič et al., 2014).

4. Annotation Platform

The dataset was divided into 400 files containing 10 tweets each and uploaded to WebAnno [1] (Eckart de Castilho et al., 2014), a general-purpose web-based annotation tool that enables multi-layer annotation. An example of annotations in WebAnno is shown in Figure 2. Yellow labels represent normalisation, green labels tokenisation, and purple labels sentence segmentation. We use special symbols to mark the deletion of a token ($0) and the end of the sentence ($.). A layer can also have multiple values (marked by "|") if a single input token should be split into more, or one word normalised into several.

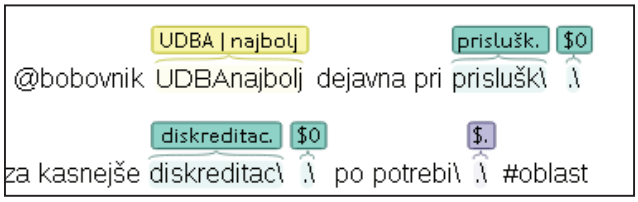

Figure 2: Annotations in WebAnno.

WebAnno was customised to allow for text annotations on the three layers relevant to our dataset: sentence segmentation, tokenisation, and normalisation.
If the same data is annotated by multiple annotators, the platform also offers a refereeing function, which enables a referee to compare multiple annotations in the same file and choose their final version.
The project workflow was designed to include a group of annotators and a referee with in-depth understanding of the annotation guidelines. The referee, who also managed the annotation campaign, designated a number of WebAnno files to each annotator group on a weekly basis. The end of each annotation phase was followed by a refereeing phase, during which the referee checked the annotations and, if necessary, provided constructive feedback to improve annotator performance by eliminating the most common and/or serious mistakes. If a particularly problematic issue arose during annotation, the annotation guidelines (see Section 5) were updated accordingly. The process was then repeated.

[1] https://webanno.github.io/webanno/

5. Annotation Guidelines

Based on a manual analysis of a small development set containing 200 randomly sampled tweets from all four categories in the dataset, annotation guidelines that address technical and linguistic aspects of the annotation process were prepared.

The technical guidelines covered the WebAnno annotation scheme and general aspects of working with the platform (joining or splitting tokens, deleting irrelevant and automatically generated tweets, dealing with complex multi-layer annotations, etc.), while the linguistic guidelines explained the criteria to follow when making language-related annotation decisions. The linguistic guidelines are summarized in the subsections below.

5.1 Sentence Segmentation

When determining sentence segmentation in tweets, the main criterion to consider is sentence-final punctuation (e.g. full stop, exclamation or question marks, two, three or multiple dots, quotes). However, tweets contain several other elements that may either appear next to sentence-final punctuation or, in its absence, fulfil a similar role. These elements are:

a) emoticons or emojis (;) =D ☺)
b) hashtags (#justsayin)
c) mentions (@author), and
d) URLs (http://t.co/fqVqV92mzc).

In the absence of sentence-final punctuation, these elements can effectively end a sentence. If the sentence ends with a series of elements, the final element is considered the end of the sentence[2]:

> Liverpool zasluženo owna Twitter, ampak na vrhu je pa fucking Iago Aspas hahaha :) #nogomet #LFC #SOULIV http://t.co/LCyEvyoVD7 ¶

If appearing after a sentence-final punctuation mark, these elements (or a series thereof) form an independent sentence:

> Življenje Je Cirkus. js sm pa čefur. Luka Stigl js sm se poscal v hlače k sm se vidu. bolano. ¶ :) ... ¶ http://t.co/QyzKRZqZnS ¶

5.2 Tokenisation

A number of elements were incorrectly split by the tokeniser and required corrections. These included abbreviations, emoticons, suffixes, and words including punctuation marks.

With abbreviations (*slov*. for *Slovene*), the tokeniser often interpreted the full stop as sentence-final punctuation and treated it as a separate token. In this case, the full stop needed to be joined with the abbreviation.

Emoticons often appeared in multiples with no spaces between and were commonly split by the tokeniser. Rather than treating each emoticon as an individual element, we decided to join the series into a single token:

- :) ¶ :) ¶ : ¶ * ¶ * → :):):**
- \ ¶ m ¶ / ¶ (¶ - ¶ _ ¶ - ¶) → \m/(-_-)

The same was done with suffixes as well as words that included punctuation:

- TV ¶ - ¶ ja → TV-ja
- sms ¶ - ¶ i → sms-i
- žen ¶ (¶ sk ¶) ¶ am → žen(sk)am
- politik ¶ (¶ e ¶ / ¶ o ¶) → politik(e/o)

5.3 Normalisation

With normalisation, two categories of words proved to be particularly problematic: non-standard words with multiple spelling variants, and foreign language elements.

5.3.1. Non-Standard Words with Multiple Spelling Variants

The first category includes non-standard words with no direct standard equivalent and multiple spelling variants (e.g. *orng*, *ornk*, *oreng*, *orenk* 'very' and *fovš*, *favš*, *fouš*, *fauš*, *fowš*, 'envious' or 'incorrect'). Such words are typically only used in spoken Slovene and have no standard spelling. In such cases, the JANES Tweet subcorpus was searched with regular expressions to find all possible spelling variants. The normalised form was then determined by selecting the most frequent one (in the above cases, *ornk* and *fouš*).

5.3.2. Foreign Language Elements

The second category consisted of foreign language elements with various degrees of adaptation to the Slovene language system in terms of spelling and morphology (e.g. *updateati*, *updajtati*, *updejtati*, *apdejtati*, 'to update'). Because of Slovene morphology, normalising these with their original language forms proved problematic (e.g. *poapdejtati*, *po-apdejt-ati*, 'to update') as it would involve introducing artificial forms absent in real language use (e.g. *po-update-ati*).

Because of this, foreign language elements were treated according to the following criteria:

a) if the word was spelled entirely phonetically (e.g. *dankešn*, 'danke schön', *aprišiejt* 'appreciate'), it would be treated as a Slovene non-standard word with multiple spelling variants (see section 5.3.1), and

b) if the word still exhibited any foreign language characteristics (e.g. non-Slovene letters or foreign language spelling), the normalised form would be the most frequent spelling variant in the JANES tweet subcorpus among those exhibiting foreign language characteristics (e.g. *updateati*, *updajtati*, *updejtati* → *updejtati*).

[2] In this paper, the end of a sentence or the delimitation between tokens is, where relevant, represented by the paragraph symbol (¶).

5.3.3. Exceptions to Normalisation

A number of Twitter- and CMC-specific elements such as mentions, hashtags, URLs, emoticons and emojis were exempt from normalisation and left in their original forms regardless of their (in)correctness.

In addition, normalisation did not extend to correcting syntactic mistakes (e.g. incorrect use of cases or mistakes in agreement, even if perceived as accidental), common lexical mistakes (e.g. using *moči* 'can' instead of *morati* 'must') or issues of style and register (*rabiti* 'to need (colloquial)' vs. *potrebovati* 'to need (standard)').

6. Annotation Campaign

In this section, we provide an overview and description of the phases of the annotation campaign.

6.1 Annotator Training

A two-day workshop was held in order to recruit annotators and familiarise them with WebAnno and the annotation guidelines. The workshop was attended by 11 annotators, all of them MA-level students of linguistics. The workshop consisted of a theoretical introduction to WebAnno, a hands-on tutorial, a presentation of the guidelines, and a training annotation session during which the participants annotated a small number of tweets. The goal of the annotation campaign was three-fold:

a) each tweet should be correctly segmented into sentences;
b) each tweet should be correctly split into tokens; and
c) all tokens should be normalised with the form closest to their standard equivalent (without radical changes to the word form, e.g. not substituting words with their standard synonyms); if the token is unclear or ambiguous, it should be left in its non-normalised form.

After the annotation session, a discussion was held to compare the annotators' decisions and the differences between them, as well as to provide correct solutions and the reasons for them in order to try and harmonise the annotators' decisions and raise inter-annotator agreement.

6.2 Annotator Testing

The workshop was followed by a test annotation session. The annotators were divided in two groups containing 5 and 6 annotators respectively. Each group was given 100 tweets from the test set and asked to correct the automatic annotations and add original annotations if necessary.

The annotations were then manually checked by the referee, who also evaluated the annotators' performance. Based on the evaluation results, 2 unreliable annotators were excluded from further assignments, and the guidelines were updated with several annotation issues that arose during the test session.

6.3 Annotation Phases and Annotator Performance

The annotation campaign was carried out in weekly phases from December 2015 to February 2016. The referee in charge of the campaign designated a number of WebAnno files to each group on a weekly basis. The remaining pool of annotators was divided into 3 groups consisting of 3 annotators.[3] A mailing list was created to allow annotators to ask questions and discuss problematic or borderline cases not included in the guidelines.

Annotator performance was monitored by measuring the annotators' effectiveness, i.e. the ratio between their annotation time and the number of tweets annotated (see Figure 3).

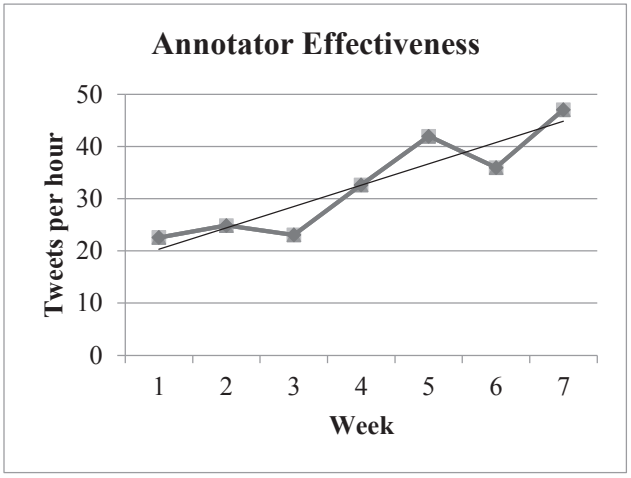

Figure 3: Annotator Effectiveness.

This was used to keep track of the annotators' weekly performance in order to optimize the flow of the annotation campaign. During the first three weeks, the annotators worked with non-standard tweets (T3L3) with a norm of 100 tweets per annotator per week. As the annotators grew more effective and the tweets steadily less noisy (T1L3, T3L1, and T1L1), the workload was increased to 150, 200, and finally 250 tweets per annotator per week. As can be deduced from Figure 2, in the case of Slovene, well-trained annotators can be expected to annotate approximately 35–45 non-standard tweets per hour – a significant improvement over the initial 21 tweets per hour.

The campaign took 7 weeks to finish, with a total of 272 hours invested by the annotators and 45 hours by the referee. On average, the annotators spent approximately 4.5 hours for each annotation session, and 30 hours for the entire campaign.

[3] When the annotators became more acquainted with the guidelines and three-member groups proved to be redundant, this number was reduced to 2, or, in the case of one accurate annotator, 1.

7. Results and Discussion

In Table 1, we give an overview of the amount of annotated tweets by standardness levels and overall. Of the initial sample of 4,000 tweets, 60 were discarded as irrelevant. In the rest, almost 10,000 sentences were identified, containing over 100,000 manually verified tokens or just under 86,600 words. It is noteworthy that the L3 tweets contain about 15% more words compared to the L1 ones. Overall, almost 12,000 words were normalised (14%), with T3L3 featuring a significantly higher number of normalisations (47%) than T1L1 (7%). The last two rows give the number of multiword normalisations, with either several original words being normalised to one word (e.g. *kvazi socializem* → *kvazisocializem*, *mega piksli* → *megapiksli*) or vice-versa (e.g. *nažalost* → *na žalost*, *nevem* → *ne vem*). The data shows that the latter category is far more frequent and also depends on the standardness level (unlike the first category).

	T1L1	T3L1	T1L3	T3L3	Total
Tweets	986	971	994	989	**3 ,940**
Sentences	2 ,413	2 ,009	2 ,934	2 ,620	**9 ,976**
Tokens	24 ,512	23 ,468	27 ,851	26 ,873	**102 ,704**
Words	20 ,333	20 ,190	22 ,912	23 ,159	**86 ,594**
Normalised words	887	1 ,136	4 ,251	5 ,570	**11 ,844**
Original multiwords	15	15	14	14	**58**
Normalised multiwords	27	63	109	139	**338**

Table 1: Quantitative Analysis of the Dataset.

During refereeing, a number of common sources of discrepancies between annotators arose. We provide a brief overview of the key problematic points for each layer in the following subsections.

7.1 Ambiguous Sentence Endings

In sentence segmentation, annotators were often faced with ambiguous sentence endings. The first category involves the use of two or multiple dots, as seen below:

> hah.. nvem ... to je pa čist odvisno od dneva ... hehe :)

The annotation guidelines required the annotators to interpret this ambiguous use of multiple dots either as a pause (which should be part of the sentence) or as sentence-final punctuation (which should end the sentence).

Similarly, in some cases, full stops, commonly used as sentence-final punctuation, were used in positions where a comma or space would be more appropriate, as seen below:

> @author1 . @author2 . @author3 . niti slucajno! kdo bo pa to placu?

A third category, especially in T3, included sentences that contained no sentence-final punctuation, but some other sign of sentence delimitation (e.g. a capital letter):

> Ko sm pa vidu to stran sm biu pa res vesel Čeprov ponavad nism za take fore :)

In many such cases, multiple (correct) interpretations were possible, which led to annotator disagreement. The final decision depended on the interpretation of the referee.

7.2 Words with Multiple Disambiguation Options

Annotators also faced ambiguity with normalisation. The most common example is the colloquial Slovene conjunction '*k*', which can be normalised to '*ko'* (when), '*ker'* (because), '*ki'* (which), and, more rarely, into '*kot'* (as) or '*kjer'* (where). The annotators were told to normalise '*k*' with the equivalent best suiting the context if possible, or to leave it in its non-normalised form if the interpretation was unclear.

A similar dilemma was posed by the word '*sm'*, which can be interpreted as either '*sem'* (I am), '*sem'* (here), or '*samo'* (only). Especially in short tweets, in which context was lacking, disambiguation proved difficult.

7.2 Misspelt Foreign Language Elements

Discrepancies between annotators were also frequent in the case of misspelt foreign language elements. According to the annotation guidelines, if a word exhibits characteristics of foreign language spelling, it should be normalised into the most frequent form exhibiting foreign language characteristics. If the word is completely foreign, it is normalised into its standard foreign language form. In the case of misspelt words like *lptop* (*laptop* vs. *leptop*) and *rter* (*router* vs. *ruter*), the annotators had to interpret the word either as foreign or as Slovene, most often by relying on the context.

7.3 Words of Ambiguous Origin

Several Slovene words, especially those containing the consonant cluster '*ks'* (*seks*, *indeks*) were often spelt using the foreign letter 'x' (*sex*, *index*). According to the annotation guidelines, Slovene words containing foreign letters should be normalised into the standard equivalents (e.g. *sex* → *seks*). Some annotators, however, interpreted these words as foreign words and left them unnormalised.

8. Conclusion

In this paper, we presented the dataset, annotation guidelines, and annotation campaign for the creation of a training dataset to be used normalisation, tokenisation, and sentence segmentation of Slovene tweets. In addition, we highlighted some of the more problematic annotation aspects which should be carefully considered when dealing with noisy social media text.

The next step in our annotation campaign will include expanding the annotated dataset with two other layers: morphosyntactic descriptions (fine grained PoS tags) and

lemmas. We will also further extend the dataset to other social media text types, in particular forum posts and on-line comments.

The latest version of the annotation guidelines (in Slovene) is available at http://nl.ijs.si/janes/viri, and the annotated dataset will be made available via the CLARIN.SI language resource repository under the Creative Commons licence (CC BY-SA 4.0). The annotation guidelines have already been adapted for Croatian and Serbian, and similar annotation campaigns are currently on-going within the ReLDI project.[4] This will allow for a cross-lingual comparison of the datasets and their impact on tagging accuracy.

9. Acknowledgements

The authors would like to thank Špela Arhar Holdt, Kaja Dobrovoljc, Simon Krek, and Katja Zupan for their valuable contributions to the annotation guidelines, as well as the annotators who participated in this project: Teja Goli, Melanija Kožar, Vesna Koželj, Polona Logar, Klara Lubej, Barbara Omahen, Eneja Osrajnik, Predrag Petrović, Polona Polc, Aleksandra Rajković, and Iza Škrjanec.

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Resources, Tools and Methods for the Research of Non-Standard Internet Slovene" (J6-6842, 2014–2017), and the Common Language Resources and Technology Infrastructure of Slovenia (CLARIN.SI).

10. References

Alegria, I., Aranberri, N., Comas, P. R., Fresno, V., Gamallo, P., Padró, L., San Vicente, I., Turmo, J., and Zubiaga, A. (2014). TweetNorm es Corpus: an Annotated Corpus for Spanish Microtext Normalization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC2014)*. ELRA, Reykjavik-Paris.

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and Wang, L. (2013). How Noisy Social Media Text, How Diffrnt Social Media Sources. In *Sixth International Joint Conference on NLP*, pp. 356–364.

Clark, E., and Araki, K. (2011). Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. In *Procedia - Social and Behavioral Sciences 27*, pp. 2–11.

Crystal, D. (2011). *Internet Linguistics: A Student Guide*. New York: Routledge.

Eckart de Castilho, R., Biemann, C., Gurevych, I. and Yimam, S. M. (2014). WebAnno: a flexible, web-based annotation tool for CLARIN. In *Proceedings of the CLARIN Annual Conference (CAC) 2014*. Soesterberg, Netherlands.

Eisenstein, J. (2013). What to Do About Bad Language on the Internet. In *NAACL-HLT*. ACL, pp. 359–369.

Erjavec, T., and Fišer, D. (2013). Jezik slovenskih tvitov: korpusna raziskava. In *Družbena funkcijskost jezika: (vidiki, merila, opredelitve), Obdobja 32*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 109–116.

Erjavec, T., Ignat, C., Pouliquen, B., and Steinberger, R. (2005). Massive multi-lingual corpus compilation: Acquis Communautaire and totale. In *Proceedings of the 2nd Language & Technology Conference, April 21–23, 2005*. Poznan, Poland, pp. 32–36.

Fišer, D., Ljubešić, N., and Erjavec, T. (2015). The JANES corpus of Slovene user generated content: construction and annotation. In *International Research Days: Social Media and CMC Corpora for the eHumanities: Book of Abstracts, 23–24 October 2015*. Rennes, France, p. 11.

Han, B., Cook, P., and Baldwin, T. (2012). Automatically Constructing a Normalisation Dictionary for Microblogs. In *EMNLP-CoNLL 2012*. Jeju, Republic of Korea, pp. 421–432.

Ljubešić, N., Erjavec, T., and Fišer, D. (2014). Standardizing tweets with character-level machine translation. In *Computational linguistics and intelligent text processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6–12, 2014: proceedings: part II, (Lecture notes in computer science, 8404)*. Heidelberg: Springer, 8404, pp. 164–175.

Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S. and Škrjanec, I. (2015). Predicting the level of text standardness in user-generated content. In *10th International Conference on Recent Advances in Natural Language Processing: Proceedings of RANLP 2015 Conference, 7–9 September 2015*. Hissar, Bulgaria, pp. 371–378.

Porta, J., and Sancho, J.-L. (2013). Word Normalization in Twitter Using Finite-state Transducers. In: *Tweet-Norm@SEPLN, Volume 1086 of CEUR Workshop Proceedings*. CEUR-WS.org, pp. 49–53.

Richard Sproat. 2001. Normalization of Non-Standard Words. In *Computer Speech & Language, 15(3)*, pp. 287–333.

Sidarenka, U., Scheffler, T., and Stede, M. (2013). Rule-Based Normalization of German Twitter Messages. In: *Proceedings of the GSCL Workshop Verarbeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation*, 2013.

Yang, Y., and Eisenstein, J. (2013). A Log-Linear Model for Unsupervised Text Normalization. In *EMNLP 2013*. ACL, pp. 61–72.

Zwitter Vitez, A., and Fišer, D. (2015). From mouth to keyboard: the place of non-canonical written and spoken structures in lexicography. In *Electronic lexicography in the 21st century*: *linking lexical data in the digital age*: *proceedings of eLex 2015 Conference*, *11-13 August 2015*, *Herstmonceux Castle*, *United Kingdom*. Ljubljana: Trojina, Institute for Applied Slovene Studies, Brighton: Lexical Computing, pp. 250–267.

[4] https://reldi.spur.uzh.ch/