THE COMPILATION, PROCESSING AND ANALYSIS OF THE JANES CORPUS OF SLOVENE USER-GENERATED CONTENT

Darja Fišer[1,2], Tomaž Erjavec[2], Nikola Ljubešić[2,3]
[1]Faculty of Arts, University of Ljubljana, Slovenia
[2]Dept. of Knowledge Technologies, Jožef Stefan Institute, Slovenia
[3]Faculty of Humanities and Social Science, Uni. of Zagreb, Croatia

1. INTRODUCTION

Despite the fact that Slovenian is fairly well-equipped with reference and specialized corpora, none of the existing resources contain user-generated content which, as is well known, has been on the rise both in terms of volume and impact in the past decade (cf. Beißwenger et al. 2014, Chanier et al. 2014). It is therefore not surprising that, apart from a few preliminary surveys of electronic texts in Slovene (Michelizza 2008, Dobrovoljc 2012, Erjavec & Fišer 2013), Slovene netspeak has not been thoroughly researched.

The *Janes*[1] corpus (Erjavec et al. 2015), the development of which is presented in this chapter, aims to change this situation. Apart from enabling a wide range of linguistic research the corpus will serve as a dataset for the development of robust tools for processing web data, which is often written without diacritics, uses phonetic spelling with lots of slang, omits punctuation etc. This is important as it has been shown that existing tools trained on standard Slovene perform poorly on this language variety (Ljubešić et al. 2014a). The presented corpus is still under construction, so it is not yet balanced or representative and contains noisy annotations but has, as the first such resource for Slovene, already proven an invaluable resource for linguistic studies and NLP experiments.

The chapter is organised as follows: Section 2 describes the corpus typology, data source selection procedure, text harvesting, linguistic annotation and a quantitative analysis of the *Janes v0.3* corpus. Section 3 focuses on the subcorpus of tweets that contains

[1] "Janes" stands for "Jezikoslovna analiza nestandardne slovenščine" (The linguistic analysis of non-standard Slovene).

the richest set of automatically and manually annotated metadata. Section 4 presents a novel approach to assign a technical and a linguistic standardness level to each text in the corpus. The paper ends with concluding remarks and plans for future development of the corpus.

2. THE JANES CORPUS

2.1. TEXT SELECTION AND CRAWLING

The current version of the *Janes* corpus contains four types of public user-generated content: tweets, forum posts, news comments and blogs. Many other popular social media, such as Facebook, Snapchat and WhatsApp contain mostly private communication, where the legal and technical barriers for the harvesting at a large scale are numerous, and were therefore not included in the corpus.

Tweets were harvested with TweetCat (Ljubešić et al. 2014b), a custom-built tool to collect tweets written in smaller languages. First, a small (about 50) set of Slovene seed words was chosen, i.e. high-frequency words which are distinct for Slovene. Using this set of seed words, accounts with predominantly Slovene posts were identified and iteratively expanded with other accounts in their social network. The tool has now been harvesting tweets for over two years, allowing us to add new tweets to the corpus at regular intervals. In addition to the content of the tweets, the tool records the associated metadata: the author's username, date and time of posting, number of retweets and likes, and the geo-tag, if available.

Due to financial and time constraints of the project we were only able to include a limited number of forums and news portals in the corpus. We chose three representative data sources for each of these two text types that generate a lot of content and/or have a large number of users. An additional factor were the publication policies and technical constraints of the providers who often lock content to be accessible to registered users only or delete it after some time.

A designated crawler and text extractor was built for each of these six text sources since they are all structured differently; the time investment needed for developing each extractor was the biggest bottleneck of the data collection process.

The forums included in the corpus comprise one of the oldest and best-known Slovene forums called *med.over.net*[2] that started as a forum for medical issues but soon spread to other topics, such as parenting, school, free time etc., and two specialized forums with a narrower focus and a more profiled user base; *avtomobilizem.com*[3] about buying, selling, servicing and enjoying cars, and *kvarkadabra.net*[4], a virtual meeting place for science enthusiasts. In addition to the content we also kept a record of the thread topic, post URL, ID, date and time, and username of its author.

News comments were obtained from the national TV portal[5], an on-line portal of the main left-wing weekly magazine *Mladina*[6] and its right-wing counterpart *Reporter*[7]. The metadata recorded with the comments are the article headline, URL, post ID, date and time of posting, and account username.

The current corpus contains a temporary collection of blogs that was extracted from the Slovene web corpus slWaC 2.0 (Erjavec & Ljubešić 2014) by taking all the documents containing the string *blog* in its URL. While the collection is already a valuable resource, a serious drawback remains the lack of separation between the blog entries and the readers' comments, as well as the lack of bloggers' metadata. This will be improved with a designated blog extractor tool that will be developed for the next version of the corpus.

The subcorpora are available separately with their complete metadata, as well as combined into a single corpus *Janes v0.3* which contains only common metadata. Encoding currently follows simple and (sub)corpus specific XML schemas.

2.2. LINGUISTIC ANNOTATION

The collected texts were tokenized and sentence segmented with a slightly modified version of the standard mlToken tokeniser for Slovene (Erjavec et al. 2005). The biggest challenge at this level remains the omission of whitespaces before/after punctuation, e.g. in "*salomon.si je zaščitena blagovna znamka"* the web domain

[2] http://med.over.net/forum5/
[3] http://www.avtomobilizem.com/forum/index.php
[4] http://forum.kvarkadabra.net
[5] http://www.rtvslo.si
[6] http://www.mladina.si
[7] http://www.reporter.si

"*salomon.si*" should be annotated as a single token, while in "*Virantova briljantna ideja.Zelo liberalno."* the string "*ideja.Zelo*" which simply lacks a space after the period, should be three tokens, including a sentence boundary. This will be tackled in the next version of the corpus when we will use a manually annotated dataset with validated token and sentence boundaries to machine-learn context-dependent sentence- and token-splitting.

Next, non-standard words in the corpus were normalized with an approach using character-based statistical machine translation and was trained on 1,000 manually normalized keywords from the tweet subcorpus with respect to a corpus of standard Slovene (Ljubešić et al. 2014a). The paired words in the lexicon were split into characters, and a standard statistical machine translation system was trained on these pairs, but instead of learning to translate sentences made up of words, the system learnt to translate words made of up characters. Even though this approach, based on translating individual words, cannot handle all non-standard phenomena in user-generated content, such as changes in tokenisation or word-order, it nevertheless produces very useful results.

Finally, the texts were morphosyntactically annotated and lemmatized using the ToTaLe tool (Erjavec et al. 2005), one of the standard corpus annotation suites for Slovene. The evaluation of lemmatization accuracy (Ljubešić et al. 2014a), which is also an implicit evaluation of morphosyntactic tagging as the assigned tag will heavily influence the quality of the lemmatization, shows that lemmatization accuracy of raw words in tweets is 75%, accuracy on manually normalized words 92%, and 84% on automatically normalized ones, which means that automatic normalization decreases the lemmatization error by half.

The corpus and its subcorpora were uploaded to our installation of the noSketch Engine web concordancer (Erjavec 2013). Access is currently restricted to project members but a publically available version of the corpus that will not infringe copyright, private information or terms of use will be released at the end of the project.

2.3. CORPUS COMPOSITION

Table 1 shows the composition of the corpus in terms of the number of words, texts, and authors per data source. All in all, there are about 135 million words and almost 5 million texts in *Janes v0.3*.

Not unusual for user-generated content, the texts are quite short, containing 28 words on average. As expected, the longest are blogs with just over 500 words per text, and the shortest are tweets, which are limited to 140 characters per message by the platform. It is interesting, however, that there is no significant difference between the average length of forum posts (50 words) and news comments (42) where one would expect forum posts to be longer. A more detailed examination of the data reveals that there are substantial differences among individual forums: the average length of posts in the medical forum is 95 words, which is almost three times more than in the automobile forum. Substantial differences are also observed among news comments: those from the national TV portal are half the size of those from the left-wing weekly.

Type	Words	Texts	Authors
Janes v0.3	134,543,613	4,819,558	385,429
tweet	50,148,724	3,684,909	7,590
forum	39,576,432	772,953	63,543
comment	12,542,551	299,420	14,295
blog	32,275,906	62,276	-

Table 1. Composition and size of *Janes* subcorpora

Texts in the corpus were contributed by more than 85,000 authors if we consider one username as one author. This is of course an estimate as the same person can use several accounts and therefore have different usernames. On average, a single author has written slightly over 1,500 words or 56 texts with the figures varying a lot among the data sources. Twitter users, for example, pen as many as eight times more texts using four times more words than the average. On the other hand, news commentators post only half the average of texts regardless of the news portal. The most deviations are found among forum posts where a user on the automobile forum posts about 18 times more texts than a user on the medical forum. On the other hand, posts on the science forum are almost twice as long as the average, falling just behind Twitter users in terms of the number of words published.

Figure 1 shows that the documents in the corpus were posted in the period 2001–2015 but almost half (49%) of them were posted in

2014 when the harvesting procedure was initiated. The oldest texts come from forums which seem to be stable enough to still record posts from as far back as 2001. The oldest news comments are from 2008 but a large majority were posted in 2014, where technical characteristics and editorial decisions of news portals play an important role. With the oldest posts from 2011, Twitter is the most recent data source. Here too, however, the most texts were collected from 2013 and 2014 coinciding with the collecting procedure. The fluctuations in 2014 are not a consequence of low traffic on Twitter but the result of technical issues with our harvesting tool.

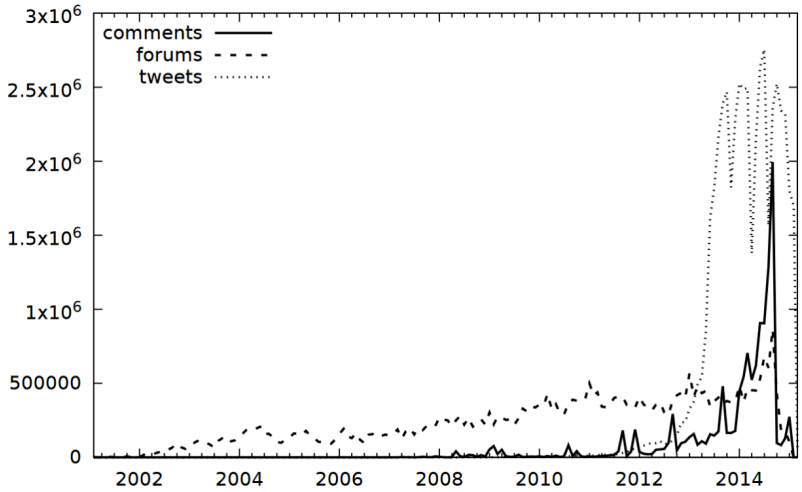

Figure 1. Age of texts (as the number of words per month) in three subcorpora of *Janes v0.3*

All this shows that the constructed corpus is very heterogeneous in terms of the authorship base, amount of texts as well as text length and age.

3. THE TWITTER SUBCORPUS

In this section we focus on the subcorpus *Janes Tweets v0.3.4* which has been updated with 500,000 additional tweets or 6 million words harvested until 23 June 2015 and enhanced with rich metadata. Table 2 shows the size and composition of this corpus. It

contains over 56 million words or 4 million tweets that were posted by about 7,600 users.

		Words	Tweets	Authors
Janes Tweet v0.3.4		58,311,996	4,337,767	7,570
Sex	female	15,417,064	1,151,300	1,858
	male	32,718,987	2,399,365	4,011
	neutral	10,175,945	787,102	1,701
Source	corporate	11,629,005	908,454	1,782
	private	46,682,991	3,429,313	5,788
Senti.	negative	15,964,247	1,006,123	-
	neutral	31,424,765	2,455,801	-
	positive	10,922,984	875,843	-

Table 2. Size and composition of the Twitter corpus

3.1. USER TYPE AND GENDER

Sociodemographic metadata is indispensable in most detailed corpus studies. Since gender in Slovene is explicitly marked in first-person past and future verb forms, we automatically identified the predominant form for each Twitter user in the corpus and then manually verified the automatic suggestions. Apart from *male* and *female* we used the *neutral* label for all accounts for which the gender of the account holder could not be determined. As shown in Table 2, men have contributed the largest share of tweets (53%) and words (56%). There are approximately half as many women (25%) who have posted 27% tweets and words. Gender could not be determined for the 22% of the accounts, which have, interestingly, contributed the lowest share of the tweets (17%) and tokens (18%). This suggests that men and women tweet with a similar frequency and length while neutral users tweet less but compose slightly longer tweets.

We also manually labelled the type of the users. Individuals who use the account for private purposes were labelled as *private* whereas accounts of news agencies, public institutions, companies, political parties etc. who use the account professionally were assigned the *corporate* label, cf. Table 2. About three quarters of the users are private (76%) while a quarter of them tweet on behalf of their company or institution (24%). Private users have contributed 79% of

the tweets or 80% of the tokens while corporate users have posted 21% of the tweets or 20% of the tokens, showing that private users tweet more and post slightly longer messages than the corporate ones. It is also interesting that, contrary to our expectations, the gender of corporate users can be determined for 20% of the accounts, 268 (80%) of which are male and 67 (20%) female.

3.2. SENTIMENT

Sentiment analysis (*positive*, *negative*, *neutral*) of user-generated content, especially tweets, is becoming increasingly popular (Pak & Paroubek 2010) as it can help better understand the opinion of the general public on a certain issue (e.g. presidential candidate, product) as well as observe trends over a period of time. We used the tool that was developed by Smailović et al. (2014), in which a Support-Vector Machine was trained on a large collection of manually annotated Slovene tweets on various topics.

Sentiment annotation was evaluated on a double-annotated sample of 1,977 tweets from the domains of *Sports* and *Politics*. Inter-annotator agreement was 75%, which shows that the task is far from trivial and quite subjective, especially in cynical and sarcastic tweets commenting on political events. If all tweets were assigned the majority sentiment, accuracy would be 37.7%, which represents a baseline system for sentiment annotation. The accuracy of the sentiment tool is 57.3% at its lowest when compared to annotator 1 and 62.1% when measured only on texts where both annotators agree on the sentiment score. These numbers are somewhat worse than necessary because the automatic approach assigns a neutral sentiment to more tweets than human annotators do, however, this makes sense from an application point of view.

Table 3 shows examples of Tweets from *Sports* and *Politics* where the automatic and manual sentiment annotations differ. In *Politics*, many of the incorrectly annotated tweets are sarcastic, cynical or ironic, which was correctly identified by the human annotators but is a well-known limitation of most sentiment annotation systems. In *Sports*, the results were better as the sentiment of the tweets was more straightforward. Where discrepancies appeared, they can often be explained by the vocabulary that is typically related to one sentiment (e.g. *victory*) but was used in a neutral, objective, factual context or in jokes. A more thorough

evaluation of the sentiment annotation of the Janes corpus can be found in (Fišer et al. 2016).

Domain	Tweet	Translation	A	M	Note
Sports	*Kot pravi vladar, čuti dolžnost, da se pred božičem pokaže ljudstvu.*	*As a true sovereign, he feels the duty to show off in front of his people before Christmas.*	0	+	sarcasm
Sports	*Slovenska RKC, daj prodaj nepremičnine in izkupiček nameni pomoči potrebnim. Za otroke gre!*	*Slovene RCC, sell your real estate and donate the profits to those in need. It's for the children!*	+	-	sarcasm
Politics	*Ekipa RD Koper 2013 zmagala na uvodnem turnirju v rokometu na mivki*	*RD Koper 2013 wins opening tournament in beach handball*	+	0	news
Politics	*Hrvati so dobili Čehe, Srbi Fince. Če oboji zmagajo, gledamo v četrtfinalu Hrvaška - Srbija! #eurobasket2015.*	*Croatians got Czech, Serbians Finns. If they both win, we will have Croatian – Serbian quarterfinals! #eurobasket2015.*	+	0	joke

Table 3. Examples of disagreements between automatic and manual sense annotations for *Sports* and *Politics*

3.3 REGION

To enable studies of regional variation in CMC we created a dataset based on the predominant region of the 1,700 different Slovene users who have posted 130,000 geo-tagged tweets (Čibej & Ljubešić 2015). Since we assume that dialects are not often used in professional communication, we only took into account private users. We divided the users into 7 traditional dialectal regions as well as

two biggest cities, the largest educational and economic centres in the country which therefore presumably act as melting pots for speakers of various dialects. We also reserved a category for tweets from abroad. In order to obtain a clean training dataset, we only considered users who have posted at least three tweets and have tweeted from the same region over 90% of the time.

These strict criteria were met by 364 users, most of which come from Ljubljana (32%) and have contributed almost 100,000 tokens to the corpus. The region with the most (44%) tweets written in non-standard language is the Alpine Gorenjsko north of Ljubljana.

4. ANNOTATING TEXT STANDARDNESS LEVEL

Preliminary analyses of the corpus had shown that the corpus contains many texts written in quite standard language. In order to be able to focus our analyses and tool development on user-generated content written in non-standard (slang, dialect etc.) Slovene, we developed an approach that assigns to each text in the corpus a standardness measure (Ljubešić et al. 2015).

We differentiate between two standardness levels: *technical* (T) and *linguistic* (L). At the technical level we observe capitalization, punctuation and spacing, which often express more the mode of entering the text (e.g. on a smart phone) than linguistic factors. The linguistic level refers to lexical choice, spelling, morphology and word order, which are the more or less concious decisions by the author to make the text more colloquial, or simply stem from their lack of awareness of the standard. Both levels have grades 1-3 where 1 means very standard and 3 very non-standard, so, for example, a T3L1 text is technically very non-standard while linguistically quite standard.

The approach is based on supervised machine learning. First, 1,200 tweets, news comments and forum posts were manually annotated. We then determined features that most likely signal text standardness for the two dimensions. Using this information, a regressor was trained, which is then able to assigns a score 1-3 for both T and L to a particular text. Evaluation of the best-performing model on a held-out test set showed that mean absolute error is T = 0.38 and L = 0.42, suggesting that technical standardness is easier to determine automatically.

Two thirds of the corpus texts are technically as well as linguistically completely standard. About a quarter of the corpus is moderately non-standard, and only 7% very non-standard. The fewest T1 and the most T3 are found in the automobile forum (45% vs. 18%) while the science forum has the largest share of T1 and the smallest T3 ones (80% vs. 3%).

On the linguistic level, there is an even higher proportion of L1 texts in the corpus (70%), 23% of them are L2 and only 7% are L3. As opposed to the technical level, the most L1 texts are found among tweets (74%) and the fewest among forum posts (50%), especially in the automobile forum (43%), which also contains the highest proportion L3 texts (19%). It is interesting to note that the proportion of T3 news comments is much larger (9%) than that of L3 (4%), most likely due to the short and formulaic nature of the comments.

5. CONCLUSION

This chapter described the collection, annotation and analysis of the first corpus of user-generated content for Slovene *Janes v0.3* as well as the *Janes Tweet v0.3.4* subcorpus. The novelty in the corpus preprocessing toolchain is the normalization of non-standard words in the corpus prior to morphosyntactic tagging and lemmatization. Another original contribution is the automatic assignment of a technical and linguistic standardness measure to the texts in the corpus that enables more focused linguistic research as well as the development of tools for the processing of non-standard Slovene. In addition, the subcorpus of tweets was annotated with valuable metadata, such as the type, gender and region of the Twitter user and the sentiment of tweets.

In our future work we plan to refine and extend the corpus with custom-extracted blogs and their comments and with Wikipedia talk pages. We will also evaluate the reliability of the developed methods for automatic linguistic and metadata annotation in order to fine-tune them for Internet Slovene. We are already working on rediacritisation of user-generated content (Ljubešić et al. 2016) and on a new part-of-speech tagger and lemmatiser (Ljubešić & Erjavec 2016). In addition, we are preparing a manually annotated dataset of tweets that will contain 4,000 tokenized, sentence-split and normalized tweets which will serve as a training and testing dataset for machine learning experiments. A major goal of the project is also

to develop a sampled and filtered *Janes* subcorpus, without copyright, private information and terms of use restrictions in order to be able to release it as open-source.

ACKNOWLEDGMENTS

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 Nonstandard Internet Slovene" (J6-6842, 2014-2017).

BIBLIOGRAPHIC REFERENCES

Beißwenger, Michael; Oostdijk, Nelleke; Storrer, Angelika & van den Heuvel, Henk (2014) : Building and Annotating Corpora of Computer-Mediated Communication: Issues and Challenges at the Interface of Corpus and Computational Linguistics. *Journal of Language Technology and Computational Linguistics.* 29(2). http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf

Chanier, Thierry; Poudat, Céline; Sagot, Benoit; Antoniadis, Georges & Wigham, Ciara R. (2014) : The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal for Language Technology and Computational Linguistics*. 29(2). 1-30. http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf

Čibej, Jaka & Ljubešić, Nikola (2015) : "S kje pa si?" : metapodatki o regionalni pripadnosti uporabnikov družbenega omrežja Twitter. In D. Fišer (ed.). *Slovenščina na spletu in v novih medijih*, Ljubljana, November 25-27 2015. Ljubljana: Znanstvena založba Filozofske fakultete, 2015, 10-14, http://nl.ijs.si/janes/wp content/uploads/2015/11/Konferenca2015.pdf.

Dobrovoljc, Helena (2008) : Jezik v e-poštnih sporočilih in vprašanja sodobne normativistike. In M. Košuta (ed.): *Slovenščina med kulturami, Zbornik Slavističnega društva Slovenije 19*. Celovec/Ljubljana: Slavistično društvo Slovenije. 295–314.

Erjavec, Tomaž (2013) : Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0*, 1/1 24–49.

http://www.trojina.org/slovenscina2.0/arhiv/2013/1/Slo2.0_2013_1_03.pdf.

Erjavec, Tomaž, Fišer, Darja & Ljubešić, Nikola (2015) : Razvoj korpusa slovenskih spletnih uporabniških vsebin Janes. *Zbornik konference Slovenščina na spletu in v novih medijih*. Ljubljana: Znanstvena založba Filozofske fakultete, 20–26. http://nl.ijs.si/janes/wp-content/uploads/2015/11/JANES15-04-Razvoj-korpusa.pdf

Erjavec, Tomaž & Fišer, Darja (2013) : Jezik slovenskih tvitov: korpusna raziskava. In A. Žele (ed.): *Družbena funkcijskost jezika (vidiki, merila, opredelitve), Obdobja 32*. Ljubljana: Znanstvena založba Filozofske fakultete. 109–116. http://www.centerslo.net/files/file/simpozij/simp32/zbornik/Erjavec.pdf

Erjavec, Tomaž, Ignat, Camelia, Pouliquen, Bruno & Steinberger, Ralf (2005) : Massive multi-lingual corpus compilation: Acquis Communautaire and ToTaLe. *Archives of Control Sciences*, 15. 529–540.

Erjavec, Tomaž & Ljubešić, Nikola (2014): The slWaC 2.0 corpus of the Slovene web. *Jezikovne tehnologije : zbornik 17. mednarodne multikonference Informacijska družba – IS 2014, October 9 –10 2014, Ljubljana,* Jožef Stefan Institute. 50–55. http://is.ijs.si/zborniki/2014_IS_CP_Volume-G_%28LT%29.pdf.

Fišer, Darja, Smailović, Jasmina, Erjavec, Tomaž, Mozetič, Igor & Grčar, Miha (2016) : Sentiment Annotation of the Janes Corpus of Slovene User-Generated Content. *Proceedings of the 10th Language Technologies and Digital Humanities Conference*, September 29-October 1 2016, Ljubljana, Faculty of Arts.

Ljubešić, Nikola, Erjavec, Tomaž & Fišer, Darja (2014a) : Standardizing tweets with character-level machine translation. *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, ISSN 0302-9743, 8404). Springer. 164–175.

Ljubešić, Nikola & Erjavec, Tomaž (2016) : Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene.

Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, 23-28 May 2016, Portorož, Slovenia.

Ljubešić, Nikola, Erjavec, Tomaž & Fišer, Darja (2016) : Corpus-based diacritic restoration for South Slavic languages. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 23-28 May 2016, Portorož, Slovenia.

Ljubešić, Nikola, Fišer, Darja & Erjavec, Tomaž, (2014b) : TweetCaT: a tool for building Twitter corpora of smaller language. *Ninth International Conference on Language Resources and Evaluation*, May 26-31, 2014, Reykjavik, Iceland. *LREC 2014*. ELRA. 2279–2283. http://www.lrec-conf.org/proceedings/lrec2014/pdf/834_Paper.pdf.

Ljubešić, Nikola, Fišer, Darja, Erjavec, Tomaž, Čibej, Jaka, Marko, Dafne, Pollak, Senja & Škrjanec, Izza (2015) : Predicting the level of text standardness in user-generated content. In *Proceedings of the International conference Recent Advances in Natural Language Processing*, Hissar, Bulgaria, 7-9 September, 2015. Hissar, 371-378, http://lml.bas.bg/ranlp2015/docs/RANLP_main.pdf.

Michelizza, Mija (2008): Jezik SMS-jev in SMS-komunikacija. *Jezikoslovni zapiski* 14/1. 151–166.

Pak, Alexander & Paroubek, Patrick (2010) : Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, 17-23 May 2010, Valletta, Malta.

Smailović, Jasmina, Grčar, Miha, Lavrač, Nada, & Žnidaršč, Martin (2014) : Stream-based active learning for sentiment analysis in the financial domain. *Information sciences*, 285:181–203.