Gold-Standard Datasets for Annotation of Slovene Computer-Mediated Communication

Tomaž Erjavec[1], Jaka Čibej[2], Špela Arhar Holdt[2,3], Nikola Ljubešić[1,4], and Darja Fišer[2,1]

[1] Department of Knowledge Technologies, Jožef Stefan Institute,
Jamova cesta 3, SI-1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si, nikola.ljubesic@ijs.si
[2] Dept. of Translation, Faculty of Arts, University of Ljubljana,
Aškerčeva cesta 2, SI-1000 Ljubljana, Slovenia
jaka.cibej@ff.uni-lj.si, spela.arharholdt@ff.uni-lj.si,
darja.fiser@ff.uni-lj.si
[3] Trojina, Institute for Applied Slovene Studies,
Trg republike 3, SI-1000 Ljubljana, Slovenia
[4] Department of Information and Communication Sciences, University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb, Croatia

Abstract. This paper presents the first publicly available, manually annotated gold-standard datasets for the annotation of Slovene Computer-Mediated Communication. In this type of language, diacritics, punctuation and spaces are often omitted, and phonetic spelling and slang words frequently used, which considerably deteriorates the performance of text processing tools that were trained on standard Slovene. Janes-Norm, which contains 7,816 texts or 184,766 tokens, is a gold-standard dataset for tokenisation, sentence segmentation and word normalisation, whereas Janes-Tag, comprising 2,958 texts or 75,276 tokens, was created for training and evaluating morphosyntactic tagging and lemmatisation tools for non-standard Slovene.

Key words: Slovene language, Computer-Mediated Communication, Word Normalisation, Morphosyntactic Tagging, Lemmatisation

1 Introduction

The development of language technologies for individual languages needs hand annotated datasets for evaluation and, with machine learning methods currantly being the dominant paradigm, also for training language models for all the relevant levels of text annotation. At least for basic text annotation, tools and datasets have already been developed for Slovene: morphosyntactic tagging and lemmatisation can be performed with ToTaLe [5] and Obeliks [11], while new tools can be trained on the openly available manually annotated corpus ssj500k [13] and the morphological lexicon Sloleks [3].

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 29–40, 2016.

However, these tools and resources predominantly deal with standard Slovene. In recent years the growing importance and quantity of Computer-Mediated Communication (CMC), such as contained in tweets and blogs, has led to a sharp increase in interest in processing such language. Tools for annotating standard language perform poorly on CMC [10], as diacritics and punctuation are often omitted, and phonetic spelling and slang words frequently used, leading to many unknown words for standard models.

The Janes[5] project aims to change this situation by developing a corpus of Slovene CMC, performing linguistic analysis on it and developing robust tools and hand-annotated gold-standard datasets for tool training and testing. It is the last goal that is the topic of this paper, which is structured as follows: Section 2 introduces the Janes corpus of Slovene CMC with an emphasis on the tools that were used to annotate it with linguistic information; Section 3 details the annotation campaign in which samples from the Janes corpus were manually annotated; Section 4 overviews the encoding, distribution and quantitative data on the resulting two datasets; and Section 5 gives some conclusions and directions for further research.

2 The Janes corpus and its annotation

The Janes corpus of Slovene CMC has been prepared in several iterations, with the current version being Janes 0.4 [8]. It contains five types of public CMC text types: tweets, forums, user comments on internet news articles (and, for completeness, also the news articles themselves), talk pages from Wikipedia and blog articles with user comments on these blogs. The collection of tweets and Wikipedia talk pages is comprehensive in the sense that the corpus includes all the users and their posts that we identified at the time of the collection. For the other texts types we selected, due to time and financial constraints, only a small set of sources that are the most popular in Slovenia and offer the most texts. Version 0.4 contains just over 9 million texts with about 200 million tokens, of which 107 come from tweets, 47 from forum posts, 34 from blogs and their comments, 15 from news comments and 5 from Wikipedia.

The texts in the corpus are structured according to the text types they belong to, e.g. conversation threads in forums, and contain rich metadata, which have been added manually (e.g. whether the author of a tweet or blog is male or female, whether the account is corporate or private) or automatically (e.g. text sentiment). For this paper, the most relevant piece of text metadata is the assignment of standardness scores to each text. We developed a method [15] to automatically classify a texts into three levels of technical and linguistic standardness. Technical standardness (T1, quite standard – T3, very non-standard) relates to the use of spaces, punctuation, capitalisation and similar, while linguistic standardness (L1 – L3) takes into account the level of adherence

[5] "Janes" stands for "Jezikoslovna analiza nestandardne slovenščine" (Linguistic Analysis of Non-Standard Slovene). The home page of the project is `http://nl.ijs.si/janes/` and the project lasts 2014–2017.

to the written norm and more or less conscious decisions to use non-standard language, involving spelling, lexis, morphology, and word order. On the basis of a manually labelled test set the method has a mean error rate of 0.45 for technical and 0.54 for linguistic standardness prediction.

The texts in the corpus have been linguistically annotated with automatic methods for five basic levels, which we describe in the remainder of this section. The tools have been developed mostly in the scope of the Janes project and typically rely on supervised machine learning. For each tool we briefly report on the training data used and, where available, the estimated accuracy of the tool.

2.1 Tokenisation and sentence segmentation

For tokenisation and sentence segmentation we used a new (Python) tool that currently covers Slovene, Croatian and Serbian [16]. Like most tokenisers, ours is based on manually specified rules (implemented as regular expressions) and uses language-specific lexicons with, e.g. lists of abbreviations. However, the tokeniser also supports the option to specify that the text to be processed is non-standard. In this case it uses rules that are less strict than those for standard language as well as several additional rules. An example of the former is that a full stop can end a sentence even though the following word does not begin with a capital letter or is even not separated from the full stop by a space. Nevertheless, tokens that end with a full stop and are on the list of abbreviations that do not end a sentence, e.g. `prof .` will not end a sentence. For the latter case, one of the additional regular expressions is devoted to recognising emoticons, e.g. `:-]`, `:-PPPP`, `^_^` etc.

A preliminary evaluation of the tool on tweets showed that sentence segmentation could still be significantly improved (86.3% accuracy), while tokenisation is relatively good (99.2%) taking into account that both tasks are very difficult for non-standard language.

2.2 Normalisation

Normalising non-standard word tokens to their standard form has two advantages. First, it becomes possible to search for a word without having to consider or be aware of all its variant spellings and, second, tools for standard language, such as part-of-speech taggers, can be used in further linguistic processing if they take as their input the normalised forms of words. In the Janes corpus all the word tokens have been manually examined and normalised when necessary by using a sequence of two steps.

Many CMC texts are written without using diacritics (e.g. `krizisce` → `križišče`), so we first use a dedicated tool [17] to restore them. The tool learns the rediacritisation model on a large collection of texts with diacritics paired with the same texts with diacritics removed. The evaluation showed that the tool achieves a token accuracy of 99.62% on standard texts (Wikipedia) and 99,12% on partially non-standard texts (tweets).

In the second step the rediacriticised word tokens are normalised with a method that is based on character-level statistical machine translation [14]. The goal of the normalisation is to translate words written in a non-standard form (e.g. `jest`, `jst`, `jas`, `js`) to their standard equivalent (`jaz`). The current translation model for Slovene was trained on a preliminary version of the manually normalised dataset of tweets presented in this paper, while the target (i.e. standard) language model was trained on the Kres balanced corpus of Slovene [18] and the tweets from the Janes corpus that were labelled as linguistically standard using the tool described above.

It should be noted that normalisation will, at times, also involve word-boundaries, i.e. cases where one non-standard word corresponds to two or more standard words or vice versa (e.g. `ne malo` → `nemalo`; `tamau` → `ta mali`). As will be shown, this raises a number of challenges both in the manual annotation and in the encoding of the final resource, as the mapping between the original tokens and their normalised versions (and their annotation) is no longer 1-1.

2.3 Tagging and lemmatisation

As the last step in the text annotation pipeline the normalised tokens are annotated with their morphosyntactic description (MSD) and lemma. For this we used a newly developed CRF-based tagger-lemmatiser that was trained for Slovene, Croatian and Serbian [16]. The main innovation of the tool is that it does not use its lexicon directly, as a constraint on possible MSDs of a word, but rather indirectly, as a source of features; it thus makes no distinction between known and unknown words. For Slovene the tool was trained on the already mentioned ssj500k 1.3 corpus [13] and the Sloleks 1.2 lexicon [3]. Compared to the previous best result for Slovene using the Obeliks tagger [11], the CRF tagger reduces the relative error by almost 25% achieving 94.3% on the testing set comprising the last tenth of the ssj500k corpus.

It should be noted that the MSD tagset used in Janes follows the (draft) MULTEXT-East Version 5 morphosyntactic specifications for Slovene[6], which are identical with the Version 4 specifications [6], except that they, following [1], introduce new MSDs for annotation of CMC content, in particular `Xw` (e-mails, URLs), `Xe` (emoticons and emojis), `Xh` (hashtags, e.g. `#kvadogaja`) and `Xa` (mentions, e.g. `@dfiser3`).

The lemmatisation, which is also a part of the tool, takes into account the posited MSD and the lexicon; for pairs word-form : MSD that are already in the training lexicon it simply retrieves the lemma, while for others it uses its lemmatisation model.

[6] `http://nl.ijs.si/ME/V5/msd/`

3 The annotation campaign

A detailed overview of the sampling procedure, the annotation workflow and guidelines, and format conversions is given in [21]; here we briefly summarise these points.

The texts that constitute the manually annotated datasets were obtained by sampling the Janes corpus. In the initial stage two samples were made: Kons1, which includes tweets, and Kons2, which includes forum posts and comments on blog posts and news articles. Kons1 contained 4,000 tweets, which were sampled randomly but taking into account some constraints. First, we removed tweets longer than 120 characters, as these are often truncated, and tweets posted from corporate accounts, which typically do not display characteristics of CMC language. Furthermore, we wanted to have a sample containing both fairly standard language (so that we don't disregard standard but nevertheless CMC specific language) as well as very non-standard ones. We therefore took equal numbers (1,000) of T1L1, T3L1, T1L3 and T3L3 tweets. Likewise, Kons2 also contained 4,000 texts and was sampled according to the same criteria as Kons1. Since, unlike Twitter, these platforms do not impose a text length limit, we here took into account only texts between 20 and 280 characters in length in order to ensure a comparable sample in text length for Kons1 and Kons2.

Having correct tokenisation, sentence segmentation and normalisation was considered a priority, so Kons1 and Kons2 were first annotated for these levels. In the second phase, the already corrected subsets of the two datasets were reimported into the annotation tool as Kons1-MSD and Kons2-MSD and MSDs and lemmatisation were corrected manually. In the selected subsets for this second annotation campaign we preferred non-standard texts to standard ones, as we were aware that the dataset will be rather small and thus wanted to make it maximally CMC-specific.

Our Guidelines for CMC annotation mostly followed the Guidelines for annotating standard [12] and historical [4] Slovene texts but with some modifications regarding the differences of the medium (e.g. emoticons, URLs). At the normalisation level, special emphasis was given to the treatment of non-standard words with multiple spelling variants and without a standard form (e.g. `orng`, `ornk`, `oreng`, `orenk` for `'very'`), foreign language elements (e.g. `updateati`, `updajtati`, `updejtati`, `apdejtati` for `'to update'`) and linguistic features that are not normalised (e.g. hashtags, non-standard syntax and stylistic issues). At the lemmatisation and MSD levels, guidelines were designed to deal with foreign language elements, proper names and abbreviations as well as non-standard use of cases and particles.

The annotation was performed in WebAnno [22], a general-purpose web-based annotation tool that enables e.g. multi-layer annotation and features with multiple values. However, the tool is difficult to use for correcting tokenisation (and hence all the token dependent layers), so we had to introduce multivalued features and some special symbols in order to be able to split and merge tokens and assign sentence boundaries, as illustrated in Figure 1. Here the string `-3,8.` was wrongly treated as one word by the tokeniser, and the annotator corrected

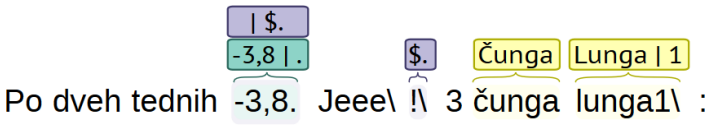

Fig. 1: Correcting token and sentence boundaries in WebAnno.

it to two tokens and inserted a sentence boundary after the second token (the full stop). It should also be noted that backslashes are used to indicate that the original text has no space between the tokens.

All the texts were first automatically annotated, then checked and corrected manually by a team of students. For the students a training and testing session was organised first. In the annotation campaigns, each text was annotated by two different annotators and then curated by the team leader.

We also put special emphasis on format conversion. The Janes corpus is encoded in TEI P5 [20], which WebAnno does not really support. We therefore developed a conversion from TEI to the WebAnno TSV tabular format, and a merge operation from the WebAnno exported TSV with the source TEI, resulting in a TEI encoding with corrected annotations. Given that we can change tokens in WebAnno this operation is fairly complex.

4 The Janes-Norm and Janes-Tag Datasets

As the end-result of the annotation we produced two datasets. Janes-Norm contains Kons1 and Kons2, i.e. it is meant as a gold-standard dataset for the annotation of tokenisation, sentence segmentation and normalisation. Janes-Tag is a subset of Janes-Norm and contains Kons1-MSD and Kons2-MSD, i.e. it is meant as a gold-standard dataset for the annotation of MSDs and lemmas. It should be noted that the order of the texts in both datasets was randomised so that it is easier to split them into training and testing sets while still retaining coverage over all text types.

4.1 Encoding and distribution

Both datasets are encoded in the same way. In particular, MSD tags and lemmas are also included in the Janes-Norm dataset, even though these were assigned automatically and thus contain errors. Nevertheless, even such annotations might prove useful for certain tasks, and it is easy enough to ignore or delete them if not needed.

Each dataset is encoded in XML as a TEI P5 [20] document, which includes its TEI header giving the metadata about the dataset and the body, which is composed of anonymous block elements (<ab>), each of which contains one text. Furthermore, each document also contains the MSD specifications encoded as a TEI feature-structure library. This makes it possible to decompose

```
<ab xml:id="janes.blog.publishwall.4264.3" type="blog" subtype="T1L3">
   <s>
      <w lemma="kaj" ana="#Rgp">Kaj</w><c> </c>
      <w lemma="biti" ana="#Va-r3s-y">ni</w><c> </c>
      <w lemma="ta" ana="#Pd-nsn">to</w><c> </c>
      <choice>
         <orig><w>tazadnje</w></orig>
         <reg>
            <w lemma="ta" ana="#Q">ta</w><c> </c>
            <w lemma="zadnji" ana="#Agpnsn">zadnje</w>
         </reg>
      </choice><c> </c>
      <choice>
         <orig><w>AAjevska</w></orig>
         <reg><w lemma="aa-jevski" ana="#Agpfsn">AA-jevska</w></reg>
      </choice><c> </c>
      <w lemma="molitev" ana="#Ncfsn">molitev</w>
      <pc ana="#Z">?</pc>
   </s>
</ab>
```

Fig. 2: TEI encoding of a text in the datasets.

an MSD into its individual features (attribute-value pairs) or localise it to Slovene.

As illustrated by Figure 2, each <ab> (i.e. a text) is labelled by its ID from the Janes corpus, its source (tweet, news, forum or blog) and its standardness score (T1L1, T1L3, T3L1 or T3L3) and then contains the contiguous sentences (<s>) containing the text. Tokens are encoded as words (<w>) or punctuation symbols (<pc>), and the original "linguistic" spacing is preserved in the TEI "character" element (<c>). Tokens are annotated with MSDs, which are pointers to their definition in the back-matter, with words also annotated with lemmas.

To encode the standard form of the words with non-standard orthography we use the TEI element <choice> with two subordinate elements, the original form(s) in <orig> and the normalised / regularised form(s) in <reg>. This complex approach has the advantage of allowing multiword mappings and distinguishing the annotation of the original from to that of the normalised form; as mentioned, we currently annotate only the normalised forms.

The TEI encoding was down-translated into the CQP vertical format used e.g. by Sketch Engine [19] and installed on the CLARIN.SI installation of noSketch Engine.

We did not perform any anonymisation on the datasets, as they are quite small and we thus do not consider them to pose a threat to privacy protection or actionable infringement of copyright or terms of use [7]. In the unlikely event

Table 1: Janes-Norm (sub)corpus sizes by standardness level and text type

	Texts		Tokens		Words		Norms		True norms		Multiw.	
All	7,816	100%	184,766	100%	143,929	100%	39,252	27.3%	16,498	42.0%	800	4.8%
T1L1	1,979	25.3%	48,438	26.2%	37,659	26.2%	7,878	20.9%	793	10.1%	78	9.8%
T1L3	1,936	24.8%	47,425	25.7%	35,569	24.7%	12,616	35.5%	6,548	51.9%	234	3.6%
T3L1	1,954	25.0%	41,474	22.4%	33,086	23.0%	6,457	19.5%	1,018	15.8%	153	15.0%
T3L3	1,947	24.9%	47,429	25.7%	37,615	26.1%	12,301	32.7%	8,139	66.2%	335	4.1%
blog	1,159	14.8%	20,987	11.4%	16,266	11.3%	3,566	21.9%	1,620	45.4%	88	5.4%
forum	1,572	20.1%	37,647	20.4%	31,002	21.5%	7,557	24.4%	3,789	50.1%	209	5.5%
news	1,145	14.6%	23,488	12.7%	19,160	13.3%	4,623	24.1%	1,876	40.6%	93	5.0%
tweet	3,940	50.4%	102,644	55.6%	77,501	53.8%	23,506	30.3%	9,213	39.2%	410	4.5%

of a complaint, we will remove the problematic text(s) from the public datasets on the concordancer(s).

We also plan to deposit Janes-Norm and Janes-Tag to the CLARIN.SI repository. Contrary to current practice in redistribution of CMC corpora (e.g. [9,2]) we will most likely distribute them under one of the CC licences.

4.2 Janes-Norm

The Janes-Norm dataset is meant for training and testing Slovene CMC tokenisers, sentence segmenters and word normalisation tools. Table 1 gives the size of the dataset overall and split into the included standardness levels and sources.

The complete dataset has 7,816 texts, which are, more or less, split equally among the four included standardness levels. The reason why the complete corpus does not have 8,000 texts and each split 2,000 texts is that the annotators had the option of marking individual texts as irrelevant (e.g. being completely in a foreign language), and these were then not included in the final dataset.

The texts contain almost 185.000 tokens or 144.000 words, where we count as a word all tokens except punctuation, numerals and tokens marked with one of the CMC-specific MSDs, i.e. emails, URLs, hashtags, mentions, emojis and emoticons. The table also shows that the proportions among the standardness levels are mostly preserved also in tokens and words. In terms of the text types, about half of the texts, tokens and words come from tweets, while about 15% of the texts are from blog and news comments each.

Moving to the number of words that have non-standard spelling, the "Norms" column shows the number of tokens that have been normalised, where the percentage is against the total number of words. "True norms" gives the number of linguistically more complex normalisations, i.e. where the normalisation goes beyond capitalisation or adding diacritics (e.g. `mačka` instead of `macka`) and the percentage refers to the number of normalised words. As can be seen, over a quarter (27.3%) of the words have been normalised, with 42% of these normalised at the morphological and lexical levels. Unsurprisingly, the

Table 2: Janes-Tag (sub)corpus sizes by standardness level and text type

	Texts		Tokens		Words		Norms		True norms		Multiw.	
All	2,958	100%	75,276	100%	56,562	100%	18,825	33.3%	10,102	53.7%	379	3.8%
T1L1	275	9.3%	6,695	8.9%	5,400	9.5%	954	17.7%	77	8.1%	11	14.3%
T1L3	1,219	41.2%	32,329	42.9%	23,159	40.9%	8,759	37.8%	4,447	50.8%	150	3.4%
T3L1	245	8.3%	4,559	6.1%	3,788	6.7%	589	15.5%	126	21.4%	12	9.5%
T3L3	1,219	41.2%	31,693	42.1%	24,215	42.8%	8,523	35.2%	5,452	64.0%	206	3.8%
blog	269	9.1%	5,046	6.7%	3,952	7.0%	848	21.5%	370	43.6%	24	6.5%
forum	403	13.6%	9,445	12.5%	7,761	13.7%	1,894	24.4%	934	49.3%	46	4.9%
news	303	10.2%	6,097	8.1%	4,801	8.5%	1,249	26.0%	522	41.8%	20	3.8%
tweet	1,983	67.0%	54,688	72.6%	40,048	70.8%	14,834	37.0%	8,276	55.8%	289	3.5%

more standard texts contain much less normalised words, with the L score significantly correlating with the need for normalisation. Looking at the text types, overall the most standard seem to be blog comments (21.9%), closely followed by forums and news comments, with tweets exhibiting the greatest proportion (30.3.%) of words requiring normalisation. The situation changes somewhat when we look at linguistically complex normalisations, as only 39.2% of tweets normalisations are the linguistically complex ones, followed by news (40.6%) and blog (45.4%) comments, and finally forums, where over half (50.1%) of the normalisations are linguistically complex, meaning that users take most care of diacritics and capitalisation on forums, and least on tweets, most likely stemming both from the instantaneous nature of the medium as well as typical input devices: forum posts on home computers vs. tweets on hand-held devices.

Finally, the last column shows the number and percentage against linguistic normalisations of cases where the normalisation involved splitting or joining words. As mentioned, these are especially difficult to model, so it is worth having a closer look at them. The results show that this is not a frequent phenomenon, involving only about 5% of linguistic normalisations. In other words, even if such cases are not treated at all, the overall drop in accuracy will not be very significant.

4.3 Janes-Tag

The Janes-Tag dataset is meant for training and testing Slovene CMC MSD taggers and lemmatisers. Similar to Janes-Norm, Table 2 gives the size of the dataset overall and split into the included standardness levels and sources.

The complete dataset has just under 3,000 texts and just over 75,000 tokens, giving over 56,000 words, i.e. it is about half the size of Janes-Norm. While this does not make for a large dataset (it is about one tenth of the size of ssj500k, the manually annotated corpus of standard Slovene) it is most likely enough to lead to significantly better Slovene CMC tagging and lemmatisation if tools were to

be trained on a combination of ssj500k and Janes-Tag. Of course, it also gives us a gold-standard dataset for testing Slovene CMC taggers and lemmatisers.

Given the sampling criteria for Kons1-MSD and Kons2-MSD the proportions of texts, tokens and words among the standardness levels is quite different from Janes-Norm, as we here concentrated on L3 texts, which make up over 80% of the dataset. In terms of text types, the majority of the texts (67%) and even more of the tokens (72.8%) come from tweets, reflecting the dynamics of the annotation campaign. The normalisation-related percentages in the table are similar to those of Janes-Norm, probably varying due to mostly random sampling factors and are here included only for the sake of completeness.

5 Conclusions

In this paper we presented two manually annotated corpora meant for training and testing tools for tokenisation, sentence segmentation, word normalisation, morphosyntactic tagging and lemmatisation of Slovene CMC. We plan to use them to improve the accuracy of tools that we have developed for these tasks, which will then be used to re-annotate the complete Janes corpus. We have also made the datasets publicly available via the concordancer and plan to make it openly available for download as well, which is highly valuable for linguistic research as no such data has so far been made available for the analysis of non-standard Slovene. In addition, it will help other researchers or companies to improve or develop their own systems for analysing this increasingly important segment of the Slovene language.

The words in the datasets have been normalised to their standard spelling but there is currently no typology of the normalisations in the dataset. And while certain types of normalisation can be easily inferred automatically (diacritisation, capitalisation, word boundaries) others cannot. In particular, phonetic spelling cannot be automatially distinguished from typos, nor can combinations of normalisation types, such as missing diacritics + phonetic spelling be recognised. This information could be useful for linguistic investigations as well as for the profiling of normalisation tools, which is why we are considering launching another annotation campaign to add this information to the normalised words in the datasets in the near future. In addition, we would find it interesting to further investigate the multiword mappings from a linguistic and technical perspectives.

Acknowledgments. The authors would like to thank 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, Dafne Marko, 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 research project "Resources, Tools and Methods for the Research of Nonstandard Internet Slovene" and the

national research programme "Knowledge Technologies", by the Ministry of Education, Science and Sport within the "CLARIN.SI" research infrastructure, and by the Swiss National Science Foundation within the "Regional Linguistic Data Initiative" SCOPES programme.

References

1. Bartz, T., Beißwenger, M., Storrer, A.: Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. Journal for Language Technology and Computational Linguistics 28(1), 157–198 (2014)
2. Chiari, I., Canzonetti, A.: Le forme della comunicazione mediata dal computer: generi, tipi e standard di annotazione. In: Garavelli, E., Suomela-Härmä, E. (eds.) Dal manoscritto al web: canali e modalità di trasmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua. Franco Cesati Editore, Firenze
3. Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M.: Morphological lexicon Sloleks 1.2. Slovenian language resource repository CLARIN.SI (2015), http://hdl.handle.net/11356/1039
4. Erjavec, T.: The IMP historical Slovene language resources. Language Resources and Evaluation pp. 1–23 (2015)
5. Erjavec, T., Ignat, C., Poliquen, B., Steinberger, R.: Massive Multilingual Corpus Compilation: Acquis Communautaire and ToTaLe. In: The 2nd Language & Technology Conference - Human Language Technologies as a Challenge for Computer Science and Linguistics. Association for Computing Machinery (ACM) and UAM Fundacja (2005)
6. Erjavec, T.: MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. Language Resources and Evaluation 46(1), 131–142 (2012), http://dx.doi.org/10.1007/s10579-011-9174-8
7. Erjavec, T., Čibej, J., Fišer, D.: Omogočanje dostopa do korpusov slovenskih spletnih besedil v luči pravnih omejitev (Overcoming Legal Limitations in Disseminating Slovene Web Corpora). Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research 4(2), 189–219 (2016), http://dx.doi.org/10.4312/slo2.0.2016.2.189-219
8. Fišer, D., Erjavec, T., Ljubešić, N.: JANES v0.4: Korpus slovenskih spletnih uporabniških vsebin (Janes v0.4: Corpus of Slovene User Generated Content). Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research 4(2), 67–99 (2016), http://dx.doi.org/10.4312/slo2.0.2016.2.67-99
9. Frey, J.C., Glaznieks, A., Stemle, E.W.: The DiDi Corpus of South Tyrolean CMC Data. In: Proceedings of the 2nd Workshop on Natural Language Processing for Computer-Mediated Communication / Social Media at GSCL2015 (NLP4CMC2015) (2015)
10. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: Annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2. pp. 42–47. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011), http://dl.acm.org/citation.cfm?id=2002736.2002747

11. Grčar, M., Krek, S., Dobrovoljc, K.: Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik (obeliks: a statistical morphosyntactic tagger and lemmatiser for slovene). In: Zbornik Osme konference Jezikovne tehnologije. Ljubljana, Slovenia (2012)
12. Holozan, P., Krek, S., Pivec, M., Rigač, S., Rozman, S., Velušček, A.: Specifikacije za učni korpus. Projekt "Sporazumevanje v slovenskem jeziku" (Specifications for the Training Corpus. The "Communication in Slovene" project). Tech. rep. (2008), http://www.slovenscina.eu/Vsebine/Sl/Kazalniki/K2.aspx
13. Krek, S., Erjavec, T., Dobrovoljc, K., Može, S., Ledinek, N., Holz, N.: Training corpus ssj500k 1.3. Slovenian language resource repository CLARIN.SI (2013), http://hdl.handle.net/11356/1029
14. Ljubešić, N., Erjavec, T., Fišer, D.: Standardizing tweets with character-level machine translation. In: Proceedings of CICLing 2014. pp. 164–75. Lecture notes in computer science, Springer, Kathmandu, Nepal (2014)
15. Ljubešić, N., Fišer, D., Erjavec, T., Čibej, J., Marko, D., Pollak, S., Škrjanec, I.: Predicting the Level of Text Standardness in User-generated Content. In: Proceedings of Recent Advances in Natural Language Processing (2015)
16. Ljubešić, N., Erjavec, T.: Corpus vs. lexicon supervision in morphosyntactic tagging: the case of slovene. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)
17. Ljubešić, N., Erjavec, T., Fišer, D.: Corpus-based diacritic restoration for south slavic languages. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA) (may 2016)
18. Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Špela., Krek, S.: Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba (The Gigafida, KRES, ccGigafida and ccKRES corpora of Slovene language: compilation, content, use). Zbirka Sporazumevanje, Trojina, zavod za uporabno slovenistiko: Fakulteta za družbene vede, Ljubljana, Slovenia (2012)
19. Rychlý, P.: Manatee/Bonito - A Modular Corpus Manager. In: 1st Workshop on Recent Advances in Slavonic Natural Language Processing. pp. 65–70. Masarykova univerzita, Brno (2007)
20. TEI Consortium (ed.): TEI P5: Guidelines for Electronic Text Encoding and Interchange.
21. Čibej, J., Špela Arhar Holdt, Erjavec, T., Fišer, D.: Razvoj učne množice za izboljšano označevanje spletnih besedil (The Developoment of a Training Dataset for Better Annotation of Web Texts). In: Erjavec, T., Fišer, D. (eds.) Proceedings of the Conference on Language Technologies and Digital Humanities. pp. 40–46. Academic Publishing Division of the Faculty of Arts, Ljubljana, Slovenia (2016), http://www.sdjt.si/jtdh-2016/en/
22. Yimam, S.M., Gurevych, I., de Castilho, R.E., Biemann, C.: Webanno: A flexible,web-based and visually supported system for distributed annotations. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013). pp. 1–6. Association for Computational Linguistics, Stroudsburg, PA, USA (Aug 2013)