Abstracts

The Janes corpus of Slovene user-generated content
Tomaž Erjavec, Nikola Ljubešić, Darja Fišer

The chapter presents the first extensive and richly annotated corpus of user-generated content for Slovene, which contains tweets, forum posts, news comments, user and talk pages from Wikipedia, and blogs and blog comments. First, we describe the harvesting procedure for each data source and provide a quantitative analysis of the corpus. Next, we present automatic and manual procedures for enriching the corpus with metadata, such as user type and gender, and text sentiment and standardness level. Finally, we present the encoding of the corpus and the procedure of making the publicly available version of the corpus, and its availability.

Keywords: corpus construction, computer-mediated communication, user-generated content, Internet Slovene, non-standard Slovene

Manually annotated Janes corpora for linguistic research and training language technology tools
Jaka Čibej, Špela Arhar Holdt, Tomaž Erjavec, Darja Fišer

In this chapter, we first present the general procedure and workflow of corpus compilation (from data preparation, guidelines, annotation platform and annotation campaign to final data conversion, publication and distribution), with particular emphasis on the largest of the corpora: Janes-Norm (approximately 185,000 tokens) and Janes-Tag (approximately 75,000 tokens), the main purpose of which is to improve language technology tools for tokenization, sentence segmentation, normalization, lemmatization and morphosyntactic tagging. The second part of the chapter consists of an overview of all manually annotated Janes corpora: in addition to the already mentioned Janes-Norm and Janes-Tag, it describes Janes-Syn (syntax in CMC), Janes-Kratko (shortening phenomena in CMC), Janes-Vejica (comma use in CMC), Janes-Preklop (code switching in CMC), and Janes-Geo (use of non-standard linguistic elements in CMC depending on the users' regional origin). The overview provides short descriptions of the content, structure and purpose of each corpus.

Keywords: Slovene, computer-mediated communication, lemmatization, normalization, morphosyntactic tagging, open data, Text Encoding Initiative, CLARIN.SI

Tools for processing non-standard Slovene
Nikola Ljubešić, Tomaž Erjavec, Darja Fišer

This chapter discusses problems associated with automatically processing non-standard language and methods we developed to resolve these problems. We consider the tasks of predicting text standardness, text segmentation, text normalisation, text rediacritisation, morphosyntactic tagging and named entity recognition. We show that the error of language

tools trained on standard language, when applied to non-standard language, increases drastically but that prior text normalisation or tool adaptation can effectively deal with non-standard input if a reasonable amount of annotated non-standard text is manually annotated and supervised machine learning models are either trained or updated on it.

Keywords: language technology, non-standard language, text normalisation, morphosyntactic tagging, named entity recognition.

Workflows for analysing non-standard Slovene
Matej Martinc, Senja Pollak, Ana Zwitter Vitez

In recent years, much effort has gone into the development of infrastructures that would simplify scientific research, increase interdisciplinary cooperation and provide easier access to research methods and data to a wide range of users. This chapter describes the implementation of a set of tools (widgets) for natural language processing into the visual programming platform ClowdFlows, which will allow linguists and other potential users to perform easier and faster text analysis. The new tools can be used for a range of different natural language processing tasks and will support corpus management and visualization of different corpus statistics. The use of the developed tools is explained and presented in two implemented workflows. The first workflow shows how the described tools can be connected into a system for building new corpora of tweets with the help of a TweetCat streaming tool. The second workflow deals with the analysis of the existing corpus of Eurovision comments, where we focus on the lexical and morphosyntactic differences between positive and negative comments. We conclude with a claim that the implemented tools enable the development of new and the analysis of existing corpora, expand the possibilities for quantitative analysis and in general reduce the complexity of natural language processing.

Keywords: natural language processing, corpus analysis tools, visual programming, ClowdFlows, non-standard Slovene

Spelling practices in Internet Slovene
Darja Fišer, Maja Miličević Petrović

This chapter presents a quantitative analysis of instances of non-standard spelling found on Slovene Twitter. The analysis is based on a manually normalized, lemmatized and part-of-speech tagged tweet sample. The focus is on transformations identified in non-standard forms (compared to the standard ones), on their distribution by part of speech and lemma, as well as the distribution of three different transformation types: deletions, insertions and replacements. The results show that a higher percentage of all transformations is covered by lexical words, but looking within PoS classes, function words are transformed to a greater extent. Deletions constitute the most common transformation

type; given that they mostly take the form of vowel deletions at word end, they point to a similarity between Slovene Twitterese and spoken language.

Keywords: non-standard spelling, computer-mediated communication, Twitter, Slovene

(Non-)standardness in Slovene CMC: the case of the comma
Damjan Popič, Darja Fišer

This chapter deals with comma placement in Slovenian tweets. We investigate to what extent comma placement in Slovenian CMC is used in compliance with standard Slovene, and in what circumstances comma placement deviates the most from the standard language. We aim to enhance previous research into the most common faults in comma placement and try to provide a more comprehensive representation of the use of the comma in Slovenian CMC, all the while making comparisons to the most recent findings in studies dealing with standard language. The results show that the standard use of the comma in the Slovenian computer-mediated communication is more common than the non-standard one. However, we can say that, in a significant portion of the dataset, the comma is omitted on purpose, in keeping with the informality of this type of communication.

Keywords: the comma, computer-mediated communication, Slovene

Regional language variants in Slovene computer-mediated communication: A corpus-based approach with the manually annotated Janes-Geo corpus
Jaka Čibej

In this chapter, we present the compilation and analysis of the manually annotated Janes-Geo corpus, which represents the first step in corpus-based studies of regional language variants in Internet Slovene. The Janes-Geo corpus contains approximately 64,000 tokens written by approximately 270 Twitter users classified into one of nine Slovene regions based on automatically generated metadata on the user's regional origin. The corpus was manually annotated with non-standard language elements according to a bottom-up typology. The purpose of the Janes-Geo corpus is two-fold: to discover the most frequent forms of linguistic non-standardness in Internet Slovene, and to compare the differences in the use of non-standard language elements between users from different regions. In addition to the method of automatically coding metadata on the regional origin of Twitter users, the chapter also describes the annotation process, the structure of the corpus, and some of the main differences between its regional subcorpora, e.g. the frequency of vowel or consonant omissions, various non-standard morphological elements, the most frequent non-standard vocabulary, and the most frequent grapheme transformations.

Keywords: regional language variants, Slovene, tweets, geolocation, computer-mediated communication

Tweets as a lexicographic resource for the analysis of semantic shifts in Slovene
Darja Fišer, Nikola Ljubešić

In this chapter we show the potential of Twitter to monitor lexicographic novelties, focusing on changes in the use of established vocabulary. The approach is based on a comparison of the target word's semantic profiles from a reference corpus and a corpus of tweets with the method of distributional modeling of words. We also propose a typology of semantically identified semantic shifts. We evaluate the results of the approach with a corpus-based manual lexicographic analysis. In addition to easily recognizable noise due to pre-processing errors in both corpora, we distinguish between, the presented approach yields valuable candidates for semantic shifts, especially those that were triggered by daily events and informal communication circumstances.

Keywords: semantic shifts, distributional semantics, corpus-based lexicography, social media

A corpus approach to syntax of computer-mediated Slovene
Špela Arhar Holdt

This chapter presents the activities that were focused on the syntax of computer-mediated Slovene. In the first part of the chapter, we describe the preparation of the Janes-Syn training corpus, a sampled corpus of 200 tweets, which were manually syntactically annotated with the JOS dependency system. We present the adaptations of the annotation system to address the following features of computer-mediated Slovene: genre-specific elements (emoticons, emojis, references to websites, user names and hashtags); the use of foreign language within the Slovene tweets; syntactical fragmentality; and nonstandard use of punctuation. The second part of the chapter presents a linguistic analysis of word order in Janes-Syn. For the study, three linguists independently annotated segments of tweets with presumably marked word order. The study examined their agreement rate, the typology of the annotated word-order problems, and the correspondence of the identified problems to the automatically assigned tags about language standardness of a specific tweet. The analysis revealed important inconsistencies in the linguistic perception of word-order markedness, while the comparison with the automatically assigned tags highlighted the need to better define the concepts of markedness and (non)standardness at the word-order level. These questions should be further addressed with the inclusion of spoken-language data. The typology of annotated problems underlined a number of previously non-examined syntactical features of computer-mediated Slovene and provided guidelines for future corpus-based research of word order in Slovene.

Keywords: Computer-mediated Slovene, corpus linguistics, syntactic annotation, tweets, word order.

Spoken elements in non-standard internet Slovene
Ana Zwitter Vitez, Darja Fišer

Communication in on-line forums, social media and news portals is frequently seen as a hybrid between spoken and written discourse. In order to examine the stereotype, we compare the features of written, spoken, and CMC discourse through an analysis of keyword forms in the corpora Kres, Gos, and Janes. The results show that at the PoS level, CMC is closer to standard written texts than spoken discourse with forum posts being the closest and tweets and news comments with a positive sentiment deviating the most from the written standard. At the lexical level, inter-speaker interactive elements that are typical of spoken discourse are most frequent in tweets and news comments. The results of the analysis could play a role in the rethinking of Slovene register variation and could also be included in the production of new language resources.

Keywords: spoken discourse, computer-mediated communication, informal communication, corpus analysis, interactive elements

The use of hashtags in Slovenian tweets
Mija Michelizza

This chapter deals with hashtags usage according to the role of hashtags in Slovenian tweets. Hashtags as a type of metadata serve as a means of categorization, but they can also perform various communication roles. Hashtags were arranged according to their role in tweets into eight categories which have been shown as relevant already in the Wikström's study (2014): topic tags, hashtag games, meta-comments, parenthetical explanations and additions, emotive usage, emphatic usage, humorous and playful usage, popular culture and tradition. It should be taken into account that categories in this categorization are not mutually exclusive; the same hashtag can be used in categorization as well as appear in any of the communication roles. In the latter, we notice that hashtags are more commonly syntactically integrated, but further research on this topic is needed. Hashtags often show the connection between tweeting while following other media as a backchannel. Some hashtags can be very long as they contain phrases or whole sentences; they rarely contain non-letter symbols other than the hash. Although hashtags represent a newer linguistic element that stands out in computer-mediated communication, they mostly remain within their roles and do not appear in most tweets from a random sample of the analyzed corpus.

Keywords: Hashtag, hash, Twitter, computer-mediated communication, Slovenian

Code-switching in Slovene tweets
Špela Reher, Darja Fišer

This chapter introduces the quantitative and qualitative analysis of code-switching (CS) in Slovenian tweets. The analysis was carried out on a sample of tweets from the Janes corpus that were manually annotated by using our own 5-level annotation scheme, which included language and type of CS, orthographic and morphologic assimilation of code-switches to the Slovenian language, as well as the part of speech. The quantitative analysis showed that CS is not a rare phenomenon, that intrasentential CS is more common than intersentential, that there were about 50% of single-word switches and that closed-class words also appear as code-switches. Among the languages used in CS, English clearly dominates, and most code-switches keep the orthographic and morphologic features of the source language. The qualitative analysis showed that CS fulfils various discourse functions, such as referential, expressive, or phatic. In terms of the semantic fields, CS was often related to popular culture, in particular TV shows, sport, food and Twitter. We also found many idiomatic expressions, phrasal verbs, collocations and even some proverbs among the code-switches.

Keywords: code-switching, borrowing, computer mediated communication, corpus linguistics

New conventions in opening and closing phrases of letters in the electronic age
Helena Dobrovoljc

The chapter looks at the old and new conventions of letter-writing, especially the opening and closing phrases. The main medium of letter communication in the past was of a physical nature, whereas today the medium is electronic. We introduce the spontaneous and learned changes in the letter discourse that are making their way into all types of electronic written communication and are an indication of how the social relationships and the writer's attitude towards letter communication have changed throughout the last century.

Key words: letter-writing, computer-mediated communication, electronic mail, opening and closing phrases, discourse elements, text

Predicting gender of Slovene bloggers
Iza Škrjanec, Nada Lavrač, Senja Pollak

Predicting the gender of text authors presents and interesting research problem; moreover, gender prediction models can be of use in various applications, such as marketing and user profiling. Our study aims to build and evaluate models for the automated gender prediction of Slovenian bloggers. For this task, we use a dataset of blog entries by 177 male and 96 female Slovenian bloggers. All blog entries by an individual user are merged and considered

a single classification instance. We compare two types of gender prediction models: a rule-based and a statistical classifier. The rule-based classification model takes into account the use of referential gender in self-referencing contexts. When building statistical models, we experiment with different features and learning algorithms. Both types of models perform with a classification accuracy over 85%. The most successful model is a token unigram learned with support vector machines. The analysis of the most informative features of this model has shown that the blogs by female and male authors display variation in terms of grammar (the use the grammatical gender and pronouns), topic (e.g. more pronounced topics of family, love, and sexuality in entries by female bloggers) and style (e.g. the use of profane language in entries by male authors).

Keywords: author profiling, gender, social media, blog classification