sciendo

Application of NLP Technologies to Low-Resource Croatian Dialects

Maja Polanec
University of Zagreb, Faculty of Electrical Engineering and Computing
Marina Bagić Babac
University of Zagreb, Faculty of Electrical Engineering and Computing

Abstract

In natural language processing (NLP) systems, a trend of decreased performance is observed when applied to texts written in low-resource dialects rather than the standard language. Dependency parsing is an essential component in NLP systems, and therefore, its improvement could lead to enhanced overall system performance. This paper aims to compare the performance of Slovenian and Croatian parsers for dependency parsing of the Kajkavian dialect. The comparison results will provide insight into the Slovenian parser's potential for parsing Kajkavian. A dependency parsing dataset was created using parallel translations of the book „Mali kraljević". Based on the created dataset, label projection from the parsed standard Croatian language to the Kajkavian dialect was performed to obtain data for calculating UAS and LAS metrics for comparing the Croatian and Slovenian parsers, which were implemented using the open-source SpaCy library. The Croatian parser achieved UAS and LAS scores of 0.47 and 0.30, respectively, which are lower than those of the Slovenian parser (0.52 and 0.34, respectively). The results indicate that the Slovenian parser performs more accurately on the Kajkavian dialect. However, to draw a general conclusion, the dataset would need to be expanded.

Keywords: Natural Language Processing (NLP); low-resource dialect; Croatian language; dependency parser

Paper type: Preliminary research

Received: Jun 14, 2024
Accepted: Jan 4, 2025

DOI: 10.2478/crdj-2025-0008

Introduction

Natural Language Processing (NLP) systems need to be sensitive to dialectal differences to ensure fairness and inclusivity for speakers of diverse dialects. However, according to Joshi et al. (2024), processing texts containing dialects tends to reduce performance in NLP systems compared to standard-language texts. Consequently, the user experience is poorer for dialect speakers, hindering equal participation of all speakers in the digital world. Qi et al. (2019) note that dependency parsing is an essential component of various NLP systems, including those for semantic role labeling, relation extraction, and machine translation. Accordingly, there is a justified motivation to explore improvements to existing dependency parsing methods, particularly for dialects.

Treebanks form the foundation for designing dependency parsers. For Croatian, there are two treebanks. The first, introduced by Tadić (2007), is HOBS (Croatian Dependency Treebank). It was created from a portion of the CW2000 subcorpus, a news corpus. The second treebank, presented by Agić and Ljubešić (2015), is Croatian UD, developed from the SETIMES: HR corpus, a collection of news articles in Balkan languages. Although, as noted by Farkaš and Filko (2022), the two Croatian treebanks differ in the procedures used for annotating specific syntactic structures during their creation, they share the common feature of being based on news corpora. All three Croatian dialects are not equally represented in newspapers, which is why the Chakavian and Kajkavian dialects can be considered low-resource dialects.

To improve dependency parsing methods for low-resource dialects, it is necessary to develop a baseline model for these dialects that serves as a reference point for comparing the performance of more advanced techniques. Baseline models can be created using zero-shot learning, where a model trained on a high-resource donor language is applied to a low-resource target language without additional adaptation. In the study by Zampieri et al. (2017), baseline models for cross-lingual dependency parsing were developed for various language pairs, including Croatian-Slovenian. Zero-shot learning was also applied in Scherrer (2014), where a Spanish parser served as a baseline for parsing Catalan. Zero-shot learning has been further explored using multiple donor languages. For example, Scherrer and Rabus (2017, 2019) trained a tagger on data from Slovak, Ukrainian, Russian, and Polish and then directly applied it to Rusyn. Zampieri et al. (2017) note that dependency parsing for Norwegian achieved better results when the model was trained on both Danish and Swedish data, compared to models trained exclusively on Danish or exclusively on Swedish.

This study aims to compare the performance of a parser trained on Croatian with that of a parser trained on Slovenian for parsing the Kajkavian dialect, to determine whether there is a difference in their potential for Kajkavian parsing. The results of such a comparison would provide insight into the feasibility of implementing a baseline Kajkavian parser using a Slovenian parser. They would help determine whether Kajkavian parsing could also benefit from Slovenian language data.

The paper is organized as follows. The chapter, Dataset for Low-Resource Dialects, provides a literature review of approaches for creating and expanding datasets for NLP processing of low-resource dialects. In the Methodology chapter, the procedure for creating the dataset is explained, the open-source library used for dependency parsing is described, and examples obtained through word alignment and label projection are presented. In the Results chapter, the Croatian and Slovenian parsers are compared using the UAS (Unlabeled Attachment Score) and LAS (Labeled Attachment Score) metrics. The Discussion chapter comments on the obtained results, provides possible explanations for these outcomes, and highlights the limitations of this study.

Dataset for Low-Resource Dialects

Pluricentric languages, such as English, are standard languages in multiple countries, with each country using its own variant. This results in the presence of these variations in numerous written sources. Consequently, when creating a dataset for further NLP processing, a large amount of material is available, including newspapers, digitized books, and more. For example, the DSLCC corpus was created from short texts extracted from newspaper articles. Among other languages in the corpus, Croatian, Serbian, and Bosnian are included. Using only the portion of this corpus that pertains to Croatian for processing Croatian dialects, without further dataset augmentation, would not be effective because all three Croatian dialects are not equally represented in the newspapers that served as the sole source for creating this dataset. Moreover, only the Štokavian dialect is represented in significant quantities. While this issue of unequal representation, or the complete absence, of particular dialects is largely absent when creating datasets for the dialects of some languages, mainly pluricentric languages, it is almost always present when processing the dialects of other languages or the colloquial speech of certain groups. Accordingly, studies on this topic present numerous methods for creating datasets for underrepresented dialects and speech varieties.

For example, in the study by Jørgensen et al. (2016), to train a model for assigning POS tags to texts associated with AAVE (African-American Vernacular English), which is a variant of English used by most working and middle-class African Americans, the training dataset was composed of Twitter posts, texts from the TV series „The Wire“, and song lyrics. The same methodology could be applied to Croatian dialects. The repertoire of songs in Chakavian and Kajkavian is sufficiently large and diverse, and these dialects also have TV series and films. However, a problem, especially with songs written in dialect, is that they are often highly stylized and focus on specific themes, which would cause models trained on such data to generalize poorly.

An alternative source for collecting data from the internet, also used in the previously mentioned study, is social media, which, as Bagić Babac (2023) notes, better reflects everyday speech than song lyrics. Alshutayri and Atwell (2017) explore this approach

for Arabic dialects, using Twitter as the data source. The authors encountered a problem also present in Croatian dialects: an uneven representation of dialects. Since the goal was to obtain a balanced dataset, they noted that running the text extractor took longer for some dialects.

Training data for dialect models is often obtained by transcribing speech from native speakers. This approach could yield the most accurate and valuable results for Chakavian and Kajkavian, as it closely reflects the language used in everyday communication. Speech transcription can be conducted using automatic speech recognition (ASR) systems. For example, Ali et al. (2016) created a corpus of an Arabic dialect using this method. Transcription can also be done manually, as in the study by Scherrer et al. (2019), which compiled a corpus of Swiss and German dialects. Created datasets can be further expanded through data augmentation. Examples of such augmentation are presented in Vania et al. (2019). They used dependency tree transformations and sentence expansion by generating new sentences. Dependency trees were transformed by removing parts of sentences, simplifying sentences, or rotating words around the root verb. Although these operations produce syntactically unnatural sentences, the authors note that this does not hinder learning and provides the model with greater variation in examples. They emphasize that this approach is particularly beneficial for languages with flexible word order and rich morphology, such as Croatian.

Methodology

Dataset creation

Most Kajkavian texts available online come from older literature and do not reflect the contemporary Kajkavian dialect. Therefore, it was not selected as a resource for creating the dataset for this study. Instead, the dataset was created using contemporary Kajkavian text. „Mali kralevič“ is a translation of the novella „Mali kraljević“ by the French author Antoine de Saint-Exupéry into modern Kajkavian. The book was translated in 2018 by Akoš Anton Dončec and Đuro Blažeka. For the purpose of creating the dataset, subtitles from the TV series „Gruntovčani“ and „Nosila je rubac črleni“were also considered, following the approach of Jørgensen et al. (2016). However, since subtitles were not publicly available at the time of writing, extracting Kajkavian from the book proved to be a more efficient and equally valuable method for creating the dataset.

Two datasets were created. One containing text from the standard Croatian version of the book, and the other from the Kajkavian version. Both datasets contain an identical number of sentences, approximately 500. The sentences in both datasets are aligned, which was crucial for applying word alignment before annotation projection. An example excerpt from the dataset, illustrating that the Croatian and Kajkavian sentences are aligned, is shown in the following table.

Table 1
Dataset overview

STANDARD CROATIAN LANGUAGE	KAJKAVIAN DIALECT
Dio 1	Del 1.
Kad mi je bilo šest godina vidio sam jednom veličanstvenu sliku, u knjizi o prašumi, koja se zvala Istinite priče.	Gda mi je bilo šest leti, videl sem jeno čudovito sliko v knigi o praega šumi koja se zove Istinite zgodbe.
Prikazivala je udava kako guta neku zvjerku.	Na sliki je bila velka kača koja je iti pôždrla neko zverino.
Evo kopije tog crteža.	Naslikal sem to sliko.
U knjizi je pisalo.	V knigi so pisali.
Udavi gutaju svoj plijen cijel cjelcat, bez žvakanja.	Velka kača pôždere svojega celoga plena, a ne vgrizne ga.
Nakon toga se više ne mogu maknuti, pa spavaju šest mjeseci, dok ga ne probave.	Onda se nemre gibati, zato prespi šest mesecov dok ga prebavla.
Zatim sam mnogo razmišljao o prašumskim pustolovinama, pa sam i sam uspio drvenom bojicom, nacrtati svoj prvi crtež.	V ovom cajto gliboko sem premišlaval o pustolovini v prašumaj i z farbanim klajbesom sem naslikal svojo prvo sliko.
Moj crtež broj 1, izgledao je ovako.	Glečte.
Pokazao sam svoje remek-djelo nekim odraslim osobama i upitao ih boje li se mojega crteža.	Pokazal sem svoje meštersko delo odraslim i pitam ak jih prestraši.

Source: Author's work

Spacy parser

According to Borotić et al. (2023), SpaCy is an open-source Python library for natural language processing. In this study, it was used for dependency parsing tasks, although it can also be applied to a range of NLP tasks such as POS tagging or morphological analysis. As noted by Šandor and Bagić Babac (2024), SpaCy supports numerous languages, including Croatian and Slovenian, whose parsers were used in this work. The models' labels in the library are compatible with the Universal Dependencies annotation scheme, enabling standardized data processing across languages. Table 2 presents the labels used by the SpaCy parser to mark dependency relations between words.

Table 2
UD parser labels

ROOT	acl	advcl	advmod	advmod:emph	amod
appos	aux	aux:pass	case	cc	ccomp
compound	conj	cop	csubj	csubj:pass	dep
det	discourse	expl:pv	fixed	flat	flat:foreign
goeswith	iobj	mark	nmod	nsubj	nsubj:pass
nummod	obj	obl	orphan	parataxis	punct

Source: Author's work

In the practical part of this study, Croatian and Slovenian dependency parsers were used. SpaCy offers models trained on both small and large datasets. For both languages, parsers trained on the most extensive datasets were used, as these parsers achieved the best performance metrics.

Word alignment

The tool SimAlign was used to align words between Croatian and Kajkavian sentences. A limitation of this approach is that it is often impossible to obtain a fully parsed sentence. Instead, the output is a parsed sentence with missing parts. An example of word alignment output for a longer sentence is shown below. From the figure, it is evident that the sentence will be incomplete when projecting labels onto the Kajkavian text, as the tool cannot align every word in the Kajkavian sentence with a corresponding word in the Croatian sentence.

Figure 1
Label projection onto Kajkavian

```
[(0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6), (7, 7), (8, 8), (9, 9), (10, 10), (11, 11),
(12, 12), (13, 13), (14, 15), (15, 16), (16, 17), (17, 18), (18, 19), (19, 20)]
```

Source: Author's work

Label projection from Croatian to Kajkavian

The Croatian parser was applied to the text „Mali kraljević“, and using the parser output along with the word alignment results, a treebank was created for the text „Mali kralevič“. When comparing label projection results with those of directly applying the Croatian parser to the Kajkavian text, it is evident that label projection yields better results. The examples below compare the results for label projection and direct parsing of Kajkavian text using a parser pre-trained on Croatian text, in the stated order.

The sentences in the examples are “Mel sem pitko vodo za komaj osem dni.“ („Imao sam pitke vode jedva za osam dana.“) i „Samo so mi rekli.“ („One su mi odgovorile.“).

```
0  Mel   _  _  _  _  0  ROOT  _  _
1  sem   _  _  _  _  0  aux   _  _
2  pitko _  _  _  _  3  amod  _  _
3  vodo  _  _  _  _  0  obj   _  _
4  komaj _  _  _  _  3  advmod _  _
5  za    _  _  _  _  7  case  _  _
6  osem  _  _  _  _  7  nummod _  _
7  dni   _  _  _  _  0  obl   _  _

0  Mel   _  _  _  _  6  nsubj _  _
1  sem   _  _  _  _  0  flat  _  _
2  pitko _  _  _  _  0  flat  _  _
```

```
3  vodo_  _  _  _  2  flat  _  _
4  za_  _  _  _  5  case_  _
5  komaj  _  _  _  _  0  nmod  _  _
6  osem  _  _  _  _  6  ROOT  _  _
7  dni  _  _  _  _  7  ROOT  _  _

0  Samo  _  _  _  _  3  nsubj_  _
1  so_  _  _  _  3  aux  _  _
2  mi_  _  _  _  3  obj  _  _
3  rekli  _  _  _  _  3  ROOT  _  _

0  Samo  _  _  _  _  1  advmod  _  _
1  so_  _  _  _  3  nsubj_  _
2  mi_  _  _  _  3  obj  _  _
3  rekli  _  _  _  _  3  ROOT  _  _
```

The examples are presented in CONLL format, which consists of three types of lines:

- Lines with words containing word annotations in 10 fields separated by tab characters. The fields, in order, are word index, the word itself, lemma, UPOS, XPOS, FEATS, HEAD, DEPREL, DEPS, MISC. Bolded fields are those filled in the CONLL files used, while the remaining fields are left empty as they are not relevant for dependency parsing.
- Empty lines that indicate sentence boundaries. The last line of each sentence is blank.
- Sentence-level comments that begin with a hash symbol (#).

Results

The metrics used to evaluate the results were UAS and LAS, which are commonly employed to assess parser model accuracy. UAS (Unlabeled Attachment Score) is calculated as the number of words correctly attached to their head divided by the total number of words in the sentence. LAS (Labeled Attachment Score) is more advanced, as it also considers the correctness of the dependency relation label. It is calculated as the number of words correctly attached to their head with the correct relation type divided by the total number of words in the sentence. For comparison, the Croatian parser from the SpaCy library achieves a UAS of 0.87 and a LAS of 0.8, while the Slovenian parser achieves 0.87 for UAS and 0.84 for LAS.

The projected labels described in the previous chapter were compared here with the outputs of the Croatian and then the Slovenian parser. Since label projection is based on word alignment, it is consequently impossible to obtain a fully parsed sentence using this method. The result is a parsed sentence with missing parts. Therefore, the outputs of the Croatian and Slovenian parsers were compared only for the words that "survived" the projection. Table 3 shows the performance of the Croatian and Slovenian parsers evaluated on approximately 500 sentences from the book „Mali kralevič".

Table 3
Results of the Croatian and Slovenian Parsers

	Croatian parser	Slovenian parser
UAS	0.47	0.52
LAS	0.30	0.34

Source: Author's work

The results show that the Slovenian parser outperforms the Croatian parser on Kajkavian texts, achieving higher UAS and LAS scores. As is commonly observed, UAS values exceed LAS values, but in this study, the difference is more pronounced than with the Croatian and Slovenian parsers from the SpaCy library. For example, for the Croatian parser in SpaCy, the difference between UAS and LAS is 0.07, whereas in this study, it reaches 0.17. For the Slovenian parser, the difference between SpaCy and this study is 0.03, while for the Slovenian parser, it is 0.18. This indicates that the model in this study has much greater difficulty assigning the correct dependency relation label than correctly attaching words to their respective heads.

Discussion

The Slovenian parser performs more accurately on the Kajkavian dialect, which can be explained by the similarities between Kajkavian and Slovenian. As noted by Celinić (2020), Kajkavian is connected to Slovenian through numerous isoglosses, and the Kajkavian dialect shares many features with Slovenian. For example, both languages exhibit the prothetic „v-„ before the reflex of „u“, and the dual, preserved in Slovenian, is partially present in Kajkavian as well. Celinić (2020) also points out that many lexical features are shared with Slovenian.

In this study, a dataset comprising approximately 500 aligned sentences in standard Croatian and Kajkavian was used. Due to the application of the label projection method, which is based on word alignment, some words in the sentences were “lost.” Such a dataset is too small to generalize the conclusion that the Slovenian parser is superior to the Croatian parser in parsing the Kajkavian dialect. Additionally, the dataset lacks diversity, as it is based solely on a single literary work.

Conclusions

This study described various approaches to creating datasets for low-resource dialects. A dependency parsing dataset was created using parallel translations of the book „Mali kraljević“. Label projection was performed from the parsed standard Croatian text to the Kajkavian dialect to obtain data for calculating UAS and LAS metrics. Finally, using these metrics, the Croatian and Slovenian parsers were compared for parsing the Kajkavian dialect. The Slovenian parser provided better results than the Croatian parser. However, due to the limitations of the dataset used, it cannot be concluded that the Slovenian parser generally performs better.

The practical significance of this research lies in the conclusion that Slovenian parsers should always be used when parsing Kajkavian, as Slovenian shares many features with Kajkavian, enabling more accurate analysis of the syntactic structure of Kajkavian texts. It would therefore make sense, when creating baseline models for testing advanced Kajkavian parsing techniques, to implement the baseline model using a Slovenian parser.

In future research, it will be necessary to expand the dataset to enable generalizable conclusions, as this study is limited by its sample size. The dataset can be expanded by creating a new dataset, and data augmentation can also be applied, as demonstrated in Vania et al. (2019).

References

1. Agić, Ž., & Ljubešić, N. (2015, September). Universal dependencies for Croatian (that work for Serbian, too). In *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing* (pp. 1–8).
2. Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., Bell, P., & Renals, S. (2016). Automatic dialect detection in Arabic broadcast speech. In *Proceedings of INTERSPEECH 2016* (pp. 2934–2938). San Francisco, CA, USA.
3. Alshutayri, A., & Atwell, E. (2017). Exploring Twitter as a source of an Arabic dialect corpus. *International Journal of Computational Linguistics (IJCL), 8*(2).
4. Bagić Babac, M. (2023). Emotion analysis of user reactions to online news. *Information Discovery and Delivery, 51*(2), 179–193. https://doi.org/10.1108/IDD-04-2022-0027
5. Borotić, G., Granoša, L., Kovačević, J., & Bagić Babac, M. (2023). Effective spam detection with machine learning. *Croatian Regional Development Journal, 3*(2), 43–64. https://doi.org/10.2478/crdj-2023-0007
6. Celinić, A. (2020). Kajkavian. *Hrvatski dijalektološki zbornik, 24*, 1–37.
7. Farkaš, D., & Filko, M. (2022). Obilježavanje koordinacije u ovisnosnim bankama stabala. *Jezikoslovlje, 23*(2), 193–214.
8. Joshi, A., Dabre, R., Kanojia, D., Li, Z., Zhan, H., Haffari, G., & Dippold, D. (2024). Natural language processing for dialects of a language: A survey. *arXiv preprint* arXiv:2401.05632. https://arxiv.org/abs/2401.05632
9. Jørgensen, A., Hovy, D., & Søgaard, A. (2016). Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1115–1120). San Diego, CA, USA.
10. Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2019). Universal dependency parsing from scratch. *arXiv preprint* arXiv:1901.10457. https://arxiv.org/abs/1901.10457

11. Scherrer, Y. (2014, August). Unsupervised adaptation of supervised part-of-speech taggers for closely related languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects* (pp. 30–38).

12. Scherrer, Y., & Rabus, A. (2017, April). Multi-source morphosyntactic tagging for spoken Rusyn. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)* (pp. 84–92).

13. Scherrer, Y., & Rabus, A. (2019). Neural morphosyntactic tagging for Rusyn. *Natural Language Engineering, 25*(5), 633–650. https://doi.org/10.1017/S1351324919000202

14. Scherrer, Y., Samardžič, T., & Glaser, E. (2019). Digitising Swiss German – How to process and study a polycentric spoken language. *Language Resources and Evaluation, 53*(4), 735–769. https://doi.org/10.1007/s10579-019-09459-5

15. Šandor, D., & Bagić Babac, M. (2024). Sarcasm detection in online comments using machine learning. *Information Discovery and Delivery, 52*(2), 213–226. https://doi.org/10.1108/IDD-01-2023-0002

16. Tadić, M. (2007). Building the Croatian dependency treebank: The initial stages. *Suvremena lingvistika, 33*(63), 85–92.

17. Vania, C., Kementchedjhieva, Y., Søgaard, A., & Lopez, A. (2019). A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 1105–1116).

18. Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., & Aepli, N. (2017, April). Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)* (pp. 1–15).

About the authors

Maja Polanec is a final year Master's student in Computer Science at the Faculty of Electrical Engineering and Computing, University of Zagreb. Two years ago, she completed her undergraduate studies in Computing at the same institution, with a thesis titled „Machine Learning Models for Music Genre Classification on the AudioSet Dataset". Her professional interests include machine learning and natural language processing. Alongside her studies, she works part-time as a programmer. She can be contacted at **maja.polanec@fer.hr**.

Marina Bagić Babac is an Associate Professor at the Faculty of Electrical Engineering and Computing, University of Zagreb, where she earned her Dipl.Ing. degree in electrical engineering, as well as her master's and PhD. She also holds a master's degree in journalism from the Faculty of Political Science at the University of Zagreb. She actively participates in several international artificial intelligence projects. She is a member of the program committees of multiple international scientific conferences and journals and serves as a reviewer for numerous international journals. Her research interests include artificial intelligence, machine learning, natural language processing, and social network analysis. She can be contacted at **marina.bagic@fer.hr**.