SUK 1.0: A New Training Corpus for Linguistic Annotation of Modern Standard Slovene

Špela Arhar Holdt[1], Jaka Čibej[1], Kaja Dobrovoljc[1,2], Tomaž Erjavec[2], Polona Gantar[1], Simon Krek[2], Tina Munda[1], Nejc Robida[1], Luka Terčon[1], Slavko Žitnik[1]

[1]University of Ljubljana, Slovenia,

[2]Jožef Stefan Institute, Ljubljana, Slovenia

{spela.arharholdt, jaka.cibej, kaja.dobrovoljc, apolonija.gantar, nejc.robida, luka.tercon}@ff.uni-lj.si

{tina.munda, slavko.zitnik}@fri.uni-lj.si

{tomaz.erjavec, simon.krek}@ijs.si

Abstract

This paper introduces the upgrade of a training corpus for linguistic annotation of modern standard Slovene. The enhancement spans both the size of the corpus and the depth of annotation layers. The revised SUK 1.0 corpus, building on its predecessor ssj500k 2.3, has doubled in size, containing over a million tokens. This expansion integrates three pre-existing open-access datasets, all of which have undergone automatic tagging and meticulous manual review across multiple annotation layers, each represented in varying proportions. These layers span tokenization, segmentation, lemmatization, MULTEXT-East morphology, Universal Dependencies, JOS-SYN syntax, semantic role labeling, named entity recognition, and the newly incorporated coreferences. The paper illustrates the annotation processes for each layer while also presenting the results of the new CLASSLA-Stanza annotation tool, trained on the SUK corpus data. As one of the fundamental language resources of modern Slovene, the SUK corpus calls for constant development, as outlined in the concluding section.

Keywords: training corpus, linguistic annotation, Slovene

1. Introduction

Training corpora play a central role in the realm of supervised machine learning for natural language processing tasks. Among these tasks, linguistic annotation stands out as a crucial step, involving the segmentation of texts into units (words, multiword units, sentences, paragraphs) and the allocation of diverse linguistic information to these units.

For machine annotation of modern standard Slovene, a training corpus has been in continuous development for more than 15 years. The previous iteration of the corpus, known as ssj500k, comprised 27,829 sentences manually labelled at multiple linguistic levels, spanning segmentation, tokenization, lemmatization, morphology and morphosyntax, dependency syntax, named entities, and semantic roles (Krek et al., 2020a).

However, the analysis of the corpus content (Arhar Holdt and Čibej, 2021) highlighted specific areas that needed refinement. As part of the Development of Slovene in a Digital Environment project,[1] the corpus underwent a significant upgrade, involving the introduction of new texts and annotations, leading to its renaming as SUK (*slovenski učni korpus*, 'the Slovene training corpus').

In this paper, we do not focus on the problems or limitations of annotating the SUK 1.0 corpus; these issues are comprehensively addressed in separate publications: Arhar Holdt et al. (2023a) and Arhar Holdt et al. (2023b) cover discussions pertaining to all annotation levels of the corpus, while Pori et al. (2022) focus on the basic levels only (lemmatization and MULTEXT-East morphosyntax), and Dobrovoljc, Terčon, and Ljubešić (2023) and Dobrovoljc and Ljubešič (2022) provide insights into Universal Dependencies aspects. Instead, the current paper serves as a concise introduction to the improved corpus for the international community. In Section 2, we compare the structure of the upgraded corpus to the previous version; in Section 3, we outline the annotation campaigns, followed by the presentation of the encoding and availability (Sections 4 and 5). The impact of the significantly expanded training corpus is demonstrated by the advancements of the CLASSLA-Stanza tagging models (Section 6).

2. Corpus Improvement

As part of the upgrade, three open-access datasets were added to the training corpus: (a) **SentiCoref 1.0** (Žitnik et al., 2022) is a corpus composed of texts from Slovene news portals published between 2007 and 2013, annotated with named entities and coreferences for sentiment analysis purposes. Its integration into SUK fulfills the demand for longer texts, enabling annotation beyond the sentence level; (b) **ELEXIS-WSD for Slovene** (Martelli et al., 2022) is the Slovene portion of a 10-language parallel corpus, containing 2,024 sentences from Wikipedia articles. It has manually annotated senses for word-sense disambiguation and, alongside SentiCoref, serves as a foundation for machine learning at the semantic level; (c) a dataset called **Ambiga** consists of 603 sentences compiled from Gigafida 2.0 (Krek et al., 2020b)—a reference corpus of standard written Slovene with texts from 1990 to 2018. It includes examples with previously underrepresented morphosyntactic tags and tokens identified as challenging for machine tagging, such as homographs and rare dual word forms.

The newly incorporated datasets were manually annotated as presented in Section 3. The expansion in scope for each annotation layer is evident from Table 1 where the data (number of tokens, sentences,

[1]Project website: https://rsdo.slovenscina.eu/en.

LREC-COLING 2024, pages 15428–15435

texts, and the corresponding percentage of the corpus) for the previous version (ssj500k) are compared to the new version of the corpus (SUK). Every layer underwent a substantial increase in size, except for verbal multiword expressions, which were not included in the project (for this layer see Krek et al., 2020a: 28). Coreferences were introduced for the first time.

Annotation layer	Tokens ssj500k I SUK	Sentences ssj500k I SUK	Texts ssj500k I SUK	% of the corpus ssj500k I SUK
segmentation	586,248 I 1,025,639	27,829 I 48,594	1,655 I 2,908	100 I 100
lemmatization, tokenization	586,248 I 1,025,639	27,829 I 48,594	1,655 I 2,908	100 I 100
MULTEXT-East morphosyntax	586,248 I 1,025,639	27,829 I 48,594	1,655 I 2,908	100 I 100
UD morphology	586,248 I 1,025,639	27,829 I 48,594	1,655 I 2,908	100 I 100
UD syntax	140,670 I 267,097	8,000 I 13,435	581 I 618	24 I 26
JOS-SYN syntax	235,864 I 267,097	11,411 I 13,435	617 I 618	40 I 26
semantic roles	112,048 I 219,216	5,501 I 11,748	228 I 598	19 I 21
named entities	194,637 I 617,832	9,488 I 29,654	498 I 1,336	33 I 60
verbal multiword expressions	280,522 I 280,522	13,511 I 13,511	754 I 754	48 I 27
coreferences	0 I 391,962	0 I 18,142	0 I 837	0 I 38

Table 1: Size of ssj500k and SUK annotation layers

3. Annotation Campaigns

The upgrade from ssj500k to SUK marks one of the most extensive annotation efforts for the Slovene language to date. Throughout the process, extensive analyses were conducted on both pre-existing annotated data and new annotation dilemmas, leading to the comprehensive enhancement of annotation guidelines across all layers.[2] A noteworthy outcome of the project is the establishment of a webpage that aggregates all updated guidelines, thus serving as a valuable resource for future work.[3]

3.1 Segmentation, Tokenization, Lemmatization and MULTEXT-East Morphosyntax

Sentence segmentation, tokenization, lemmatization, and morphosyntactic tagging using the MULTEXT-East v6 system are the fundamental corpus annotation layers, so all newly added texts have been tagged and subjected to a thorough manual review across these layers.

The first three layers were reviewed simultaneously by 9 annotators, following the principles from the Obeliks rule-based tokenizer.[4] The morphosyntax was manually revised in a separate campaign. During this phase, 24 annotators re-evaluated the automatically assigned morphosyntactic tags (or MSD-tags) over an approximate duration of four months. In the process, which was based on the established guidelines by Holozan et al. (2008), the tag of each token was assessed by three distinct annotators, following the principle of triple agreement: tags unanimously selected by all three annotators were final, while tags with discrepancies were re-examined during the curation phase (for a detailed description of the methodology, see Pori et al., 2022). Following the curation phase, the data were subjected to a series of semi-automated consistency checks. In light of the challenges and dilemmas encountered throughout the morphosyntax revision process, refinements were made to corresponding sections in the existing guidelines.

A statistical overview of the most frequent dilemmas and corrections has shown that the automatic lemmatization and morphosyntactic tagging for written standard Slovene has advanced to a stage where it would be sensible to transition from comprehensive manual revisions to more targeted ones (ibid.: 165). Nonetheless, implementing such targeted revisions necessitates the formulation of well-documented and referenced methodologies for either automatic or semi-automatic detection of contentious points.

3.2 Universal Dependencies

The format Universal Dependencies is a framework for cross-linguistically consistent annotation of grammar (parts of speech, morphological features,

[2]While the paper does not present inter-annotator agreement metrics (with the exception of Universal Dependencies, Section 3.2), SUK is an expert-curated resource. Curation, seamlessly integrated into the annotation process, was overseen by linguistics experts. Disagreements and dilemmas that emerged were not only regularly resolved but also became focal points for in-depth discussions, leading to the refinement of annotation guidelines.

[3]https://wiki.cjvt.si/shelves/linguistic-annotation-of-slovene-corpora

[4]https://github.com/clarinsi/obeliks

and syntactic dependencies) aimed at facilitating scientific advances in multilingual technology and research on language typology (de Marneffe et al., 2021). As part of the development of the SUK training corpus, the reference Slovene UD treebank (SSJ) originating from the ssj500k corpus has been substantially improved and extended to almost double the original size (Dobrovoljc and Ljubešić 2022).

The process was based on the initial revision and documentation of the language-specific UD annotation guidelines for Slovene, followed by a two-stage annotation campaign using the Q-CAT (Brank, 2023) and WebAnno (Eckart de Castilho et al., 2016) annotation tools. First, the annotators added dependency relations to the 3,411 partially parsed sentences that remained unreleased at the time of the original SSJ treebank creation (Dobrovoljc et al., 2017), followed by a manual inspection of 2,024 sentences from the automatically pre-parsed ELEXIS-WSD corpus. More details on the annotation process and the inter-annotator agreement are given by Dobrovoljc and Ljubešić (2022).

In addition to being included as part of the larger SUK corpus, the extended version of the SSJ treebank has also been released as part of UD v2.10 (Zeman et al., 2022) in CoNLL-U format.

In contrast to the dependency parsed (SSJ) subset, which only represents a quarter of the SUK corpus, morphology-related UD tags (part-of-speech categories and morphological features) have been assigned to all the tokens of the SUK corpus by semi-automatic mapping from the manually checked MULTEXT-East morphosyntactic tags (Section 3.1).[5]

3.3 JOS-SYN Dependency Syntax

The JOS-SYN system, developed within the framework of the JOS: Linguistic Annotation of Slovene project (Erjavec et al., 2010), aligns with the insights of Slovene linguistics (Toporišič, 2004) while adhering to the foundational principles of dependency syntax annotation. An essential aspect of JOS-SYN is its integration with MULTEXT-East for Slovene (Erjavec, 2012). At the syntactic layer, JOS-SYN strategically avoids duplicating information covered by morphosyntax, ensuring a robust and interpretable annotation system. This approach allows for high-precision parsing (Table 2), and intuitive utilization of the parsed data.

The annotation campaign spanned approximately four months. Initially, the ELEXIS-WSD sentences, having undergone manual corrections for tokenization, segmentation, lemmatization, and MULTEXT-East MSD-tags, were parsed using CLASSLA-Stanza (version 1.1.0). Subsequently, two annotators examined each sentence, correcting labels with the help of the Q-CAT annotation tool (Brank, 2023). While we addressed several annotation issues by updating the annotation guidelines (e.g., the improved annotation of proper names and symbols (Arhar Holdt et al., 2023a: 132–134)), certain challenges exceeding the scope of syntax, such as the treatment of foreign language elements, will need to be further addressed in future projects.

3.4 Semantic Role Labeling

Semantic Role Labeling (SRL) refers to detecting and assigning semantic roles to semantic arguments determined by the predicate or a verb in a sentence. A semantic role annotation scheme for Slovene (Gantar et al., 2018) follows the Prague Dependency Treebank tagset (Mikulová et al., 2006) and is adapted to the Slovene language specifics. The final version of the Slovene tagset consists of 25 semantic labels: five for arguments, 17 for adjuncts and three labels for multi-word predicates and other multiword expressions.

The semantic annotation process involved 11,748 sentences, of which 5,501 sentences (formerly manually annotated in the previous version of the training corpus) were reviewed by two annotators, and 6,247 sentences were first automatically annotated using the SRL parser (Björkelund, 2009) and then manually checked. The Q-CAT tool (Brank, 2023) was used for annotation. The decisions were finally synchronized across the SRL subcorpora.

We upgraded the annotation by rethinking existing decisions in line with new findings in producing semantic resources for Slovene (Gantar, 2021; Gantar, 2023). For example, in about 75% of the corpus, the relations between the arguments of the verbs describing speech act are corrected and unified according to the pattern: REC = addressee of the verb action, RESLT = concrete result or "product" of the verb action, PAT = object or topic of the verb action.

Other substantive improvements to the corpus are based on unifying decisions related to the syntactic layer. This includes defining the subject-verb agreement in copulative clauses of the following type:

(1) Gostja večera bo Desa Muck. (Slovene)
'The guest of the evening will be Desa Muck.'
(2) Pravilni odgovor je Grad Podčetrtek. (Slovene)
'The correct answer is Podčetrtek Castle.'

In Slovene, identifying who/what is the originator of the action, the bearer of the event or a quality/property (typically denoted by ACT), as well as who/what the object of the action or the action refers to (typically denoted by PAT), presents a notable challenge. To ensure maximum consistency at the semantic level, which can also serve as a foundation for decision-making at the syntactic level, we adopted a guideline that the semantic interpretation should align with the principle: what I learn new = affected participant (typically PAT), about whom or what I learn something new = originator of the action or bearer of the property (typically ACT). We have also partially unified the decisions in understanding agent and de-agent structures. In annotating, we followed the semantic interpretation of the initiator of the action (ACT), to which it is usually impossible to add another initiator without changing the meaning: *the event (ACT) took place – *ACT took place the event.*

[5]The rules and conversion scripts are available at https://github.com/clarinsi/jos2ud. The mapping is fully automatic, apart from the verb *biti* ('to be') which requires manual VERB-AUX disambiguation.

3.5 Named Entity Recognition

Named entity recognition involves identifying and classifying entities such as personal names, locations, and organizations. For creating SUK, the established guidelines were used (Zupan et al., 2023). The pre-existing labels in SentiCoref were upgraded, while Elexis-WSD was annotated anew.

A team of three annotators and a curator used the INCEpTION tool (Klie et al., 2018) end-to-end. A label for each named entity was independently evaluated by three annotators. Labels that received consensus from all three were accepted, whereas the named entities that were attributed conflicting labels, underwent additional review in the curation phase.

The dilemmas faced during the revision underscored the necessity of introducing additional categories for possessive adjectives derived from proper names labelled not only as 'person' but also as 'organization' on one hand and 'location' on the other (Arhar Holdt et al., 2023a: 140). The integration of these categories would significantly alter the current annotation system, necessitating further analyses and a careful evaluation.

3.6 Coreference Resolution

The goal of coreference resolution is to identify and link all mentions that refer to a common entity in a text. Anaphora and coreferences form a subset of discourse parsing (Soricut and Marcu, 2003), which is crucial for text understanding.

In SUK, coreferences could be annotated within the newly incorporated SentiCoref corpus as it consists of coherent and complete texts. SentiCoref was already annotated with coreferences, however, without following specific good practices. For the new annotation campaign, we adapted Serbian coreference annotation guidelines that were created within the ReLDI 2008 project.[6] These guidelines followed practices by the ACE 2004 evaluations with added specifics for Slavic languages. Compared to the ReLDI guidelines, we do not annotate syntactic features as they are available within other annotation layers. We also improved the structure and terminology of the guidelines.

Coreference resolution was annotated within the INCEpTION platform (Klie et al., 2018). Each text was annotated by two annotators and final annotations were decided by a curator. As INCEpTION does not offer curation support for the coreference resolution, the curator needed to manually compare two documents and adjust annotations in one document that was selected as the final one.

Slovene has some specifics that needed to be considered, such as referencing in negations, implicit mentions within verb forms or multiple references to an entity in a single sentence. As part of our future work, annotation decisions should be evaluated based on a state-of-the-art coreference tagger, and guidelines further improved.

4. SUK Encoding

Just as ssj500k, the canonical encoding of SUK is in XML following the TEI Guidelines,[7] however, against the upgraded TEI parameterization as recommended by the CLARIN.SI research infrastructure.[8] Because SUK is composed of several subcorpora, which contain different metadata on the contained texts and different layers of linguistic annotation, the corpus is composed of the top-level file with the TEI header giving the corpus-wide metadata, and links to the files of the subcorpora. Each of these files then contains divisions with the annotated texts.

While TEI is a very expressive annotation scheme, its use requires specialized software and familiarity with both XML and TEI. In the field of language technology, the much simpler CoNLL-U format, developed in the scope of the Universal Dependencies project, has recently become a de-facto standard, which is why we have developed a conversion procedure from our TEI to CoNLL-U and made the corpus available also in this format. However, the SUK CoNLL-U format does not encompass the more complex annotation layers, such as coreferences. Furthermore, since SUK contains syntactic dependencies in two formalisms (JOS-SYN and UD), every syntactically annotated subcorpus has two CoNLL-U files: one with the UD syntax and the MSD-tags in English, and another with JOS syntax and the MSD-tags in Slovene.

Examples of the TEI and CoNLL-U encoding of SUK are given in Appendix.

5. Availability

The SUK corpus in TEI and CoNLL-U is available from the CLARIN.SI repository under the CC BY-SA 4.0 licence (Arhar Holdt et al., 2022), i.e., it is also available for commercial exploitation. The corpus is, for browsing and analysis, also available through the CLARIN.SI concordancers noSketch Engine (Rychlý, 2007) and KonText (Machálek, 2020), with the links to the concordancers given in the repository entry.

6. Models for Linguistic Annotation

A prime example of the SUK 1.0 corpus's practical application is its utilization in the training of models for linguistic annotation of texts. The CLASSLA-Stanza[9] linguistic processing pipeline was employed to train the models, as it is particularly well adapted to the specific features of South Slavic languages (Terčon and Ljubešić, 2023). The tool has also been used to train models using the ssj500k training corpus (Ljubešić and Dobrovoljc, 2019).

Models were trained for five annotation tasks and evaluated. The performance scores for each task are shown in Table 2. The scores are summarized using the F1 score of the labels, with syntactic dependency parsing using the F1 score of the commonly employed labelled attachment score, i.e. LAS (Nivre and Fang, 2017). The positive impact of the considerably larger

[6] https://wiki.cjvt.si/books/09-coreferences/page/annotation-guidelines

[7] https://tei-c.org/release/doc/tei-p5-doc/en/html

[8] https://github.com/clarinsi/TEI-schema, see also https://www.clarin.si/repository/xmlui/page/data#tei.

[9] https://pypi.org/project/classla/

training corpus is detailed by Terčon et al. (2023), where a comparison between models trained on both corpus editions is given.

Annotation task	F1
Morphosyntactic tagging (MULTEXT-East)	97.08
Lemmatization	98.97
UD syntax	90.57
JOS-SYN syntax	93.89
Semantic role labelling	76.24

Table 2: Model performance for each annotation task

7. Conclusion

Training corpora for linguistic annotation form the fundamental digital infrastructure for any language, necessitating their continuous improvement. We presented the upgrade from ssj500k 2.3 to SUK 1.0 and demonstrated the positive impact of this substantial endeavor on the annotation of modern standard Slovene.

Future work comprises several key priorities for the further development of the SUK 1.0 corpus. Firstly, a comprehensive evaluation of the final product across all annotation levels is essential, paving the way for the continual refinement of the corpus and associated annotation guidelines across linguistic levels, addressing challenges that extend across annotation layers. The focus is on achieving methodological consistency within the corpus and in relation to other linguistic resources. Additionally, there is a need to increase the corpus size and enhance genre representation, such as by incorporating texts from the legal and academic domains. There's room to enhance semantic and discourse levels through the inclusion of new annotation types. Ensuring the ongoing development of tools for linguistic annotation, analysis, and visualization, with a user-friendly approach to accessing richly annotated data, is crucial. Finally, active coordination and participation within international annotation initiatives and on standardized datasets is important for facilitating cross-lingual methodology and aligning with broader global research efforts.

8. Acknowledgments

The authors acknowledge that the project Development of Slovene in a Digital Environment was funded by the Ministry of Culture of the Republic of Slovenia and the European Union from the European Regional Development Fund in the period 2020–2023. The programs Language Resources and Technologies for Slovene (P6-0411), Slovene Language – Basic, Contrastive, and Applied Studies (P6-0215) and Knowledge Technologies (P2-0103) received financial support from the Slovenian Research Agency. The authors are also grateful to the reviewers for their constructive remarks and to the annotators who participated in the annotation process.

9. Bibliographical References

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Erjavec, T., Gantar, P., Krek, S., Munda, T., Robida, N., Terčon, L., and Žitnik, S. (2023a). Nadgradnja učnega korpusa ssj550k v SUK 1.0. In Š. Arhar Holdt and S. Krek (Eds.), *Razvoj slovenščine v digitalnem okolju* (1. izd.). Založba Univerze, pp. 119–156. https://doi.org/10.4312/9789612972561.

Arhar Holdt, Š., Bordon, D., Čibej, J., Dobrovoljc, K., Gantar, P., Lenardič, J., Munda, T., Pori, E., Robida, N., Terčon, L., and Žitnik, S. (2023b). *Slovenski učni korpus: množici SUK 1.0 in Janes-Tag 3.0: poročilo projekta Razvoj slovenščine v digitalnem okolju: aktivnost DS1.2.* https://www.cjvt.si/rsdo/wp-content/uploads/sites/18/2023/06/RSDO_Kazalnik_SUK_v2.pdf.

Arhar Holdt, Š. and Čibej, J. (2021). Analize za nadgradnjo učnega korpusa ssj500k. In Š. Arhar Holdt (Ed.), *Nova slovnica sodobne standardne slovenščine*: *viri in metode*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 15–53.

Björkelund, A., Hafdell, L., and Nugues, P. (2009). Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task,* pp. 43–48, Boulder, Colorado, June. Association for Computational Linguistics.

De Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational linguistics*, 47(2):255–308.

Dobrovoljc, K., Erjavec, T., and Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pp. 33–38, Valencia, Spain. Association for Computational Linguistics.

Dobrovoljc, K. and Ljubešić, N. (2022). Extending the SSJ Universal Dependencies Treebank for Slovenian: Was It Worth It? In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022,* pp. 15–22, Marseille, France. European Language Resources Association.

Eckart de Castilho, R., Mujdricza-Maydt, E., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A web-based tool for the integrated annotation of semantic and syntactic structures. In *Proceedings of LT4DH*, pp. 76–84, Osaka, Japan. The COLING 2016 Organizing Committee.

Erjavec, T., Fišer, D., Krek, S., and Ledinek, N. (2010). The JOS Linguistically Tagged Corpus of Slovene. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10),* pp. 1806–1809, Valletta, Malta. European Language Resources Association.

Erjavec, T. (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):131–142.

Gantar, P., Strkalj Despot, K., Krek, S., and Ljubešić, N. (2018). Towards semantic role labeling in Slovene and Croatian. In D. Fišer and A. Pančur (Eds.). *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, pp. 93–98, Ljubljana,

Slovenija. Ljubljana: Znanstvena založba Filozofske fakultete.

Gantar, P. (2021). Strojno berljiv Vezljivostni leksikon slovenskih glagolov. In Š. Arhar Holdt (Ed.), *Nova slovnica sodobne standardne slovenščine*: *viri in metode*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 259–297.

Gantar, P. (2023). Analiza udeleženskih vlog s skladenjskega, pomenskega in leksikalnega vidika. In M. Smolej and M. Schlamberger Brezar (Eds.). *Prispevki k preučevanju slovenske skladnje*. Ljubljana: Založba Univerze, pp. 77–97.

Holozan, P., Krek, S., Pivec, M., Rigač, S., Rozman, S., and Velušček, A. (2008). *Specifikacije za učni korpus (Specifications for the training corpus). Version 1.0.* Kazalnik K2 projekta Sporazumevanje v slovenskem jeziku. Available from: https://wiki.cjvt.si/books/04-multext-east-morphosyntax/page/annotation-guidelines.

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., and Gurevych, I. (2018). The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations,* pp. 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J., and Brank, J. (2020a). The ssj500k training corpus for Slovene language processing. In D. Fišer and T. Erjavec (Eds.). *Jezikovne tehnologije in digitalna humanistika: zbornik*, pp. 24–33, Ljubljana, Slovenija. Ljubljana: Inštitut za novejšo zgodovino.

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., and Dobrovoljc, K. (2020b). Gigafida 2.0: the reference corpus of written standard Slovene. In N. Calzolari, Nicoletta (Ed.). *LREC 2020: Twelfth International Conference on Language Resources and Evaluation*, pp. 3340–3345, Marseille, France. European Language Resources Association.

Ljubešić, N. and Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pp. 29–34, Florence, Italy. Association for Computational Linguistics.

Machálek, T. (2020). KonText: Advanced and Flexible Corpus Query Interface. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 7003–7008, Marseille, France. European Language Resources Association.

Martelli, F., Navigli, R., Krek, S., Tiberius, C., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Pedersen, B. S., Olsen, S., Langements, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña-Ruiz, R.-J. ..., and Munda, T. (2021). Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. In I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, and C. Tiberius. (Eds.). *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference*, pp. 377–395, virtual. Brno: Lexical Computing CZ, s.r.o.

Mikulová, M., Bémová, A., Hajič, J. …, and Žabokrtský, Z. (2006). Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Technical Report 30. 5–11.

Nivre, J. and Fang, C. (2017). Universal Dependency Evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pp. 86–95, Gothenburg, Sweden. Association for Computational Linguistics.

Pori, E., Čibej, J., Munda, T., Terčon, L., Arhar Holdt, Š. (2022). Lematizacija in oblikoskladenjsko označevanje korpusa SentiCoref. In D. Fišer and T. Erjavec (Eds.). *Jezikovne tehnologije in digitalna humanistika: zbornik konference*, pp. 162–168, Ljubljana, Slovenija. Ljubljana: Inštitut za novejšo zgodovino.

Rychlý, P. (2007). Manatee/Bonito-A Modular Corpus Manager. In P. Sojka and A. Horák (Eds.). *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pp. 65–70, Brno. Brno: Masaryk University.

Soricut, R. and Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 228–235.

Terčon, L. and Ljubešić, N. (2023). CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. *arXiv*. https://doi.org/10.48550/arXiv.2308.04255.

Terčon, L, Ljubešić, N. and Dobrovoljc, K. (2023) Doubling the Amount of Training Data: Does It Help? A New Training Corpus for Slovene and Its Impact on Automatic UD Annotation. In *UniDive 1st general meeting - selected abstracts*. Paris, France.

Toporišič, J. (2004). *Slovenska slovnica*. Maribor: Obzorja.

Zupan, K., Ljubešić, N., and Erjavec, T (2023). *Annotation guidelines for Slovenian named entities Janes-NER. Version 1.1.* Available from: https://wiki.cjvt.si/books/08-named-entities/page/annotation-guidelines.

Žitnik, S., Blagus, N., and Bajec, M. (2022). Target-level sentiment analysis for news articles. *Knowledge-Based Systems*, 249(August 2022).

10. Language Resource References

Arhar Holdt, Š. et al., (2022). *Training corpus SUK 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1747.

Brank, J. (2023), *Q-CAT Corpus Annotation Tool 1.5*. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1844.

Zeman, D. et al. (2022). *Universal Dependencies 2.10*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-4758.

Appendix A: Example of TEI encoding

As an example of the canonical TEI encoding of SUK we give the complete annotation (except for omitting some UD morphological features) of the sentence: Ima kakšnega prijatelja? (Slovene)
'Does (s)he have a friend?'

```
<s xml:id="ssj6.32.112">
  <w ana="mte:Ggnste-n"
    msd="UPosTag=VERB|...|VerbForm=Fin"
    lemma="imeti"
    xml:id="ssj6.32.112.t1">Ima</w>
  <w ana="mte:Zv-met"
    msd="UPosTag=DET|...|PronType=Int"
    lemma="kakšen"
    xml:id="ssj6.32.112.t2">kakšnega</w>
  <w ana="mte:Sometd"
    msd="UPosTag=NOUN|...|Number=Sing"
    lemma="prijatelj"
    xml:id="ssj6.32.112.t3"
    join="right">prijatelja</w>
  <pc ana="mte:U"
     msd="UPosTag=PUNCT"
     lemma="?"
     xml:id="ssj6.32.112.t4">?</pc>
  <linkGrp corresp="#ssj6.32.112"
         targFunc="head argument"
         type="UD-SYN">
    <link ana="ud-syn:root"
        target="#ssj6.32.112 #ssj6.32.112.t1"/>
    <link ana="ud-syn:det"
        target="#ssj6.32.112.t3
#ssj6.32.112.t2"/>
    <link ana="ud-syn:obj"
        target="#ssj6.32.112.t1
#ssj6.32.112.t3"/>
    <link ana="ud-syn:punct"
        target="#ssj6.32.112.t1
#ssj6.32.112.t4"/>
  </linkGrp>
  <linkGrp corresp="#ssj6.32.112"
         targFunc="head argument"
         type="JOS-SYN">
    <link ana="jos-syn:modra"
        target="#ssj6.32.112 #ssj6.32.112.t1"/>
    <link ana="jos-syn:dol"
        target="#ssj6.32.112.t3
#ssj6.32.112.t2"/>
    <link ana="jos-syn:dve"
        target="#ssj6.32.112.t1
#ssj6.32.112.t3"/>
    <link ana="jos-syn:modra"
        target="#ssj6.32.112 #ssj6.32.112.t4"/>
  </linkGrp>
  <linkGrp corresp="#ssj6.32.112"
         targFunc="head argument"
         type="SRL">
    <link ana="srl:PAT"
        target="#ssj6.32.112.t1
#ssj6.32.112.t3"/>
  </linkGrp>
</s>
```

Appendix B: Example of CoNLL-U encoding

As an example of the derived CoNLL-U encoding we give below (Tables 3 and 4) the same sentence as for the TEI encoding, in both variants, i.e. with the UD annotations and with the JOS annotations. The examples are rendered in the form of a table for better readability. In the actual text encoding, the row breaks correspond to the newline character, while the column breaks correspond to the tab character:

# sent_id = ssj6.32.112									
# text = Ima kakšnega prijatelja?									
1	Ima	imeti	VERB	Vmpr3s-n	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Pres|VerbForm=Fin	0	root	_	NER=O
2	kakšnega	kakšen	DET	Pq-msa	Case=Acc|Gender=Masc|Number=Sing|PronType=Int	3	det	_	NER=O
3	prijatelja	prijatelj	NOUN	Ncmsay	Animacy=Anim|Case=Acc|Gender=Masc|Number=Sing	1	obj	_	NER=O|SpaceAfter=No
4	?	?	PUNCT	Z	_	1	punct	_	NER=O

Table 3: Example of the CoNLL-U encoding with UD annotations

# sent_id = ssj6.32.112									
# text = Ima kakšnega prijatelja?									
1	Ima	imeti	glagol	Ggnste-n	nikalnost=nezanikani|oblika=sedanjik|oseba=tretja|vid=nedovršni|vrsta=glavni|število=ednina	0	modra	_	NER=O
2	kakšnega	kakšen	zaimek	Zv-met	sklon=tožilnik|spol=moški|vrsta=vprašalni|število=ednina	3	dol	_	NER=O
3	prijatelja	prijatelj	samostalnik	Sometd	sklon=tožilnik|spol=moški|vrsta=občno_ime|število=ednina|živost=da	1	dve	_	NER=O|SpaceAfter=No
4	?	?	ločilo	U	_	0	modra	_	NER=O

Table 4: Example of the CoNLL-U encoding with JOS annotations