

# 12 Slovene learner corpus KOST

The KOST annotation system was developed together with the KOST corpus of Slovene as a foreign language (Stritar Kučuk 2022) and is designed for categorizing teacher's corrections in texts written by speakers of Slovene as a second or foreign language. The tagging system is hierarchically organized in two tiers: first, the corrections are defined according to the linguistic level, followed by the characterization of the general type of correction or the part of speech. The two-tier annotations allow for a robust analysis, which has to be followed by a more detailed manual revision.

- [Introduction to Tags](#)
- [Annotation Guidelines](#)
- [References and Links](#)

# Introduction to Tags

This chapter summarises the KOST tags. A more detailed presentation can be found in the guidelines in the Annotation Guidelines chapter.

| Tag     | Linguistic level | Type of correction/part of speech    |
|---------|------------------|--------------------------------------|
| Z-LOC   | orthography      | punctuation                          |
| Z-CRK   | orthography      | spelling                             |
| Z-SN    | orthography      | joined or divided words              |
| Z-MV    | orthography      | capitalization                       |
| Z-KR    | orthography      | abbreviation                         |
| B-SAM   | vocabulary       | noun                                 |
| B-GLAG  | vocabulary       | verb                                 |
| B-ZAIM  | vocabulary       | pronoun                              |
| B-PRID  | vocabulary       | adjective                            |
| B-PRISL | vocabulary       | adverb                               |
| B-PRED  | vocabulary       | preposition                          |
| B-VEZ   | vocabulary       | conjunction                          |
| B-OST   | vocabulary       | other                                |
| O-SAM   | word form        | noun                                 |
| O-GLAG  | word form        | verb                                 |
| O-ZAIM  | word form        | pronoun                              |
| O-PRID  | word form        | adjective                            |
| O-PRISL | word form        | adverb                               |
| O-OST   | word form        | other                                |
| S-STR   | syntax           | structure                            |
| S-BR    | syntax           | word order                           |
| S-IZP   | syntax           | omission                             |
| S-ODV   | syntax           | insertion                            |
| POV     | /                | related correction                   |
| [???    | /                | incomprehensible, unclear correction |

# Annotation Guidelines

This chapter summarizes the annotation guidelines for semantic-role labelling as applied to Slovene texts. The guidelines are arranged from the latest, up-to-date version to the oldest version.

## **Version 1.0 (04-2022)**

**Project** [Development of Slovene in a Digital Environment](#)

STRITAR KUČUK, Mojca, 2023: *KOST 1.0: Priročnik za označevanje napak, delovna verzija*. Različica

1.0. [\[PDF\]](#) - only in Slovene

# References and Links

This chapter compiles relevant references and provides links to projects where the KOST system has been developed and applied to Slovene texts.

## **Projects, in which the system has been developed:**

[Development of Slovene in a Digital Environment](#)

## **Corpora containing manually revised KOST tags:**

STRITAR KUČUK, Mojca, ŠTER, Helena, PISEK, Staša, PETRIC LASNIK, Ivana, KETE MATIČIČ, Jana, PIRIH SVETINA, Nataša, PREGLAU, Daniela, ARHAR HOLDT, Špela, KRSNIK, Luka, ERJAVEC, Tomaž, 2023, *Slovene learner corpus KOST 1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1753>.

## **The CJVT Svala tool for manual annotation following the KOST system:**

ARHAR HOLDT, Špela, KOSEM, Iztok, STRITAR KUČUK, Mojca, KRSNIK, Luka, JOVAN, Leon Noe, 2022: *CJVT Svala* (Kazalnik projekta Razvoj slovenščine v digitalnem okolju), v1.0, <https://orodja.cjvt.si/svala/>, dostop 2. 3. 2023.

## **References:**

STRITAR KUČUK, Mojca, 2022: *KOST med korpusi usvajanja tujega jezika*. Obdobja 41: Na stičišču svetov: slovenščina kot drugi in tuji jezik. 323–334. [https://centerslo.si/wp-content/uploads/2022/11/Stritar-Kucuk\\_Obdobja-41.pdf](https://centerslo.si/wp-content/uploads/2022/11/Stritar-Kucuk_Obdobja-41.pdf)

ARHAR HOLDT, Špela, KOSEM, Iztok, STRITAR KUČUK, Mojca, 2022: *Metode in orodja za lažjo pripravo korpusov usvajanja jezika*. Obdobja 41: Na stičišču svetov: slovenščina kot drugi in tuji jezik. 23–30. [https://centerslo.si/wp-content/uploads/2022/11/Arhar-Holdt-et-al\\_Obdobja-41.pdf](https://centerslo.si/wp-content/uploads/2022/11/Arhar-Holdt-et-al_Obdobja-41.pdf)

STRITAR KUČUK, Mojca, 2020: *Modul Leto plus – prvi korak do korpusa slovenščine kot tujega jezika*. Zbornik konference Jezikovne tehnologije in digitalna humanistika 2020. 131–135. [http://nl.ijs.si/jtdh20/pdf/JT-DH\\_2020\\_StritarKucuk\\_Modul-Leto-plus%e2%80%93prvi-korak-do-korpusa-slovenscine-kot-tujega-jezika.pdf](http://nl.ijs.si/jtdh20/pdf/JT-DH_2020_StritarKucuk_Modul-Leto-plus%e2%80%93prvi-korak-do-korpusa-slovenscine-kot-tujega-jezika.pdf)