Alternative Endings of Slovene Verbs in Third Person Plural: A Corpus Approach

Gašper Pesek, Iza Škrjanec, Dafne Marko
Ljubljana
E-mail: gasper.pesek@gmail.com, skrjanec.iza@gmail.com, dafne.marko@gmail.com

Abstract

This paper is concerned with the alternative endings of some Slovene verbs in third person plural. Certain Slovene verbs in this form can take the endings *-jo* or *-do* (*jejo* or *jedo*), whereby the two possible choices are normatively evaluated in various ways in Slovene language manuals. The paper introduces a corpus-driven approach to the problem by first extracting potential verbs with this phenomenon, after which the use of both endings is compared in the subcorpora of Janes (a corpus of Slovene user-generated content) and in the Kres corpus.

Keywords: alternative endings, Slovene verbs, user-generated content, standard language

1. Introduction

Slovene is a morphologically rich language with some alternative forms in the inflection paradigm. Related studies using a corpus approach (Može, 2013; Arhar Holdt, 2013) show that Slovene language manuals either insufficiently describe the phenomena in question, or prescriptively evaluate one variant as more prestigious than the other. This paper is concerned with Slovene verbs that can take the endings *-do* or *-jo* in third person plural. We observe their behavior in the Janes corpus of user-generated Slovene (Fišer et al., 2016) and in the Kres corpus (Logar Berginc et al., 2012), which mostly contains standard written Slovene.

The rest of the paper is structured as follows: in Section 2, the alternative verb endings *-jo* and *-do* are presented together with their evaluation in different Slovene language manuals. In Section 3, we briefly present the two corpora and the data extraction process. The verb forms and their frequencies in the Janes subcorpora and in the Kres corpus are analyzed in Section 4, while Section 5 provides a genre comparison. Section 6 concludes the paper and suggests further work.

2. Problem Description and the Aim of the Paper

Certain Slovene verbs can take two different endings in third person plural: *-jo* or *-do*. In Slovene grammar (Toporišič, 2004), this characteristic is observed in five athematic verbs (*jejo – jedo*; *grejo – gredo*; *bojo – bodo; vejo – vedo; dajo – dado*; 'they eat/go/will be/know/give', respectively). According to paragraph 891 in the Slovene normative language manual *Slovenski pravopis 2001*, the ending *-jo* is frequently used instead of *-do*, which is especially true for "literary conversational language" (*knjižni pogovorni jezik*) – considered less appropriate for written texts – and derivatives (by means of prefixation) of the previously mentioned athematic verbs, e.g. *prepovejo – povedo,* where the derivative with the prefix *pre-* is written with the ending *-jo*, whereas the original verb is written with the ending *-do*. The choice between two alternative forms seems to be puzzling for the users of Slovene, as can be seen from several questions posted in specialized[1] as well as general[2] online forums, such as Med.Over.Net. In the replies, two main factors for the use of appropriate endings are emphasized: the medium (spoken or written) and register (literary or conversational). The ending *-do* is considered more common in written language and formal communication.

In this regard, an analysis of the phenomenon in computer-mediated communication or user-generated content would be interesting, as the language on the Internet is heavily influenced by spoken language (Crystal, 2006); however, a spectrum of genres are considered as CMC, differing in synchronicity, message size, privacy settings, communication norms, etc. (for a list of factors, see Herring, 2007).

The aim of this paper is to use a corpus-based approach to determine the general tendencies of using the endings *-do* and *-jo*. We expect to find a preference for the ending *-jo* in the Janes corpus and a preference for the ending *-do* in the Kres corpus. We intend to observe the usage patterns with regard to different genres in Janes, and to contrast the tendencies in the Janes corpus with those in the Kres corpus, which is a corpus of standard written language.

3. Methodology

3.1 Corpus Description

For the analysis, two corpora were queried. The Janes v0.4 corpus is a corpus of user-generated Slovene. It contains over 175 million words or 9 million documents, published between 2002 and 2016 in five different genres: tweets, forum posts, blog entries and their comments, online news and their comments, and user and page talk from the Slovene Wikipedia.

[1] E.g.: http://www2.arnes.si/~lmarus/suss/arhiv/suss-arhiv-000103.html (accessed: 10 August, 2016)

[2] E.g.: http://med.over.net/forum5/viewtopic.php?t=2547265, http://med.over.net/forum5/viewtopic.php?t=10424263 (accessed: 10 August, 2016)

The Kres corpus, which contains nearly 100 million words, is a collection of standard written Slovene with a balanced genre structure: it consists of periodicals (newspapers and magazines), fiction and non-fiction, documents from the Web, and other genres, published between 1990 and 2011.

3.2 Data Extraction

The Sketch Engine concordance (see Kilgarriff, 2014) was used for corpus scanning. Employing a CQL expression, we used the Kres corpus to first extract verbs that end in *-do*, after which we noted the frequency of the forms with *-do* and *-jo*[3] in the five Janes subcorpora[4], as well as the entire Janes and Kres corpora. Thus, we collected 17 verbs[5]: *biti* ('be'), *iti* ('go'), *dati* ('give'), *vedeti* ('know'), *izvedeti* ('find out'), *zvedeti* ('find out'), *poizvedeti* ('inquire'), *zavedeti* ('realize'), *jesti* ('eat'), *pojesti*[6] ('eat up'), *najesti* ('sate'), *povedati* ('tell'), *izpovedati* ('confess'), *dopovedati* ('get across'), *napovedati* ('predict'), *odpovedati* ('cancel'), and *prepovedati* ('forbid').

4. Analysis

4.1 Distributions of Variants in the Janes Subcorpora

Because Slovene's 5 athematic verbs are a linguistically unique group, they call for a separate initial analysis. With regard to the future tense form of the verb *biti*, all subcorpora indicate a virtually exclusive preference (nearly 100% of concordances) for the ending *-do*. The opposite is true of *dati*, where *-jo* has almost completely replaced its older alternative[7]. In the case of *vedeti*, *-do* is preferred in all subcorpora – least prominently in the forum subcorpus, and most prominently in the blog (nearly 90% of concordances) and Wiki talk subcorpora (over 90%)[8]. *Iti* displays a relatively even distribution between the two alternatives, with the exception of the blog subcorpus, where a preference for *-do* is observed (roughly 80%). *Jesti* has revealed a general preference for the *-do* ending, with a relatively equal distribution in the tweet subcorpus, a slight preference for *-do* in the forum subcorpus, and a strong preference for *-do* in the comment (roughly 70%), blog (roughly 80%), and Wiki talk (roughly 85%) subcorpora. The following paragraphs summarize the specifics of each subcorpus.

In the tweet subcorpus (Figure 1[9]), the *-do* ending is preferred (60% or more) for *najesti* and *povedati*; *-jo* is preferred for *izpovedati*, *napovedati*, *odpovedati*, and *prepovedati*. All other verbs display a relatively equal distribution of both endings.

The forum subcorpus (Figure 2) reveals a considerable preference (roughly 70%) for the *-do* ending with the verb *zvedeti*. A relatively equal distribution of both endings is observed with *izvedeti*, *zavedeti*, *pojesti*, and *povedati*. All other verbs show a preference (60% or more) for *-jo*, especially in the case of *izpovedati* and *dopovedati*, which have only produced concordances with *-jo*.

In the blog subcorpus (Figure 3), there is an overall preference for *-do*. A relatively even distribution of both endings can be seen for *poizvedeti*, *zavedeti*, *najesti*, *izpovedati*, *odpovedati*, and *prepovedati*. *Napovedati* shows a notable preference (over 60%) for *-jo*.

The news comment subcorpus (Figure 4) reveals a notable preference (60% or more) for *-do* in the cases of *izvedeti* and *povedati*, while *zvedeti* and *poizvedeti* have only provided concordances with *-do*. *Zavedeti*, *pojesti*, and *izpovedati* have shown a relatively equal distribution of the two endings, whereas the others (*najesti*, *napovedati*, *odpovedati*, and *prepovedati*) have shown a notable preference for *-jo*. *Dopovedati* did not produce any concordances.

In Wiki talk (Figure 5), the verbs that have produced concordances show a strong preference for the *-do* ending, with the exception of *povedati* and *izvedeti*, which have displayed a relatively even distribution of the two ending variants.

4.2 The Janes Subcorpora in Relation to Kres

This subsection examines each of the Janes subcorpora in relation to Kres. The 5 athematic verbs will once again be described separately.

In all of the Janes subcorpora, the ratios of *-jo* to *-do* for the verb *biti* are virtually the same as in Kres. For *iti*, on the other hand, all Janes subcorpora display a much stronger preference for *-jo*, except for the blog subcorpus, as its ratio of *-jo* to *-do* is practically the same as in Kres. *Dati* almost exclusively employs the *-jo* ending in both Kres as well as all of the Janes subcorpora. For *vedeti*, the tweet, forum, and news comment subcorpora prefer *-jo* compared to Kres, whereas the blog and Wiki talk subcorpora show ratios similar to the ones in Kres (which shows a slightly stronger preference for *-do*). Finally, *jesti* displays a notable preference for *-jo* in the tweet, forum, and news comments subcorpora. There is an overlap with the Kres ratio in the Wiki talk subcorpus, and a slight preference for *-jo* with the ratio in the blog subcorpus.

The distributions of alternate endings for the verbs *poizvedeti*, *zavedeti*, and *dopovedati* in the tweet subcorpus are very similar to those in Kres. With *najesti*, however,

[3] CQL expression for verb extraction:
[word=".+do" & tag="G.*"]
CQL expression for form frequency, e.g., *bojo*:
[word="(b|B)ojo" & tag="G.*"]

[4] In the news subcorpus, only the comment section was taken into consideration, excluding the news because they are not user-generated content.

[5] The verb *zajesti* was also extracted from the Kres corpus, but was not included in the analysis due to low or zero frequencies in the Janes subcorpora.

[6] Because the third person plural forms of the verbs *pojesti* ('eat up') in *peti* ('sing') overlap (*pojejo* – 'they eat/sing'), all concordances were analyzed manually.

[7] Since the word form *dado* is used mostly as a proper noun (but has been erroneously annotated as a verb form), all concordances were analyzed manually.

[8] There is an overlap between the word forms *vedo* (third person plural) and *vedo* (a participle in masculine singular form). The latter is a regional, non-standard variation of the participle *vedel*, used mainly in northeastern Slovenia. Since all derivatives of the verbs *vedeti*, *jesti* and *povedati* follow the same pattern, there might be a slight deviation from the actual frequency of verbs ending in *-do* for third person plural.

[9] For charts, see Appendix 1.

there is a higher preference for *-do* in the tweets. All other verbs show a notable to strong preference for *-jo* in comparison with Kres.

In the forum subcorpus, all remaining verbs show a higher preference for *-jo*, albeit in varying degrees.

The verbs in the blog subcorpus display a slightly higher preference for *-jo* with the verb *zavedeti*, and a stronger preference for *-jo* with *zvedeti*, *pojesti*, *najesti*, and *napovedati*. The ratios for *poizvedeti*, *povedati*, and *odpovedati* are similar to the ones in Kres, while the remaining verbs (*izvedeti*, *izpovedati*, *dopovedati*, and *prepovedati*), seem to prefer *-do* more than Kres.

In the news comment subcorpus, *pojesti*, *najesti*, *povedati*, *napovedati*, and *odpovedati* show a higher preference for *-jo*. The ending ratios for *izvedeti*, *zavedeti*, *izpovedati*, and *prepovedati* are almost the same as the ones in Kres. Interestingly, *zvedeti* and *poizvedeti* show a higher preference for *-do*.

In the Wiki talk subcorpus, there is a slightly higher preference for *-jo* with *izvedeti* and a notably higher preference for *-jo* with *povedati*. *Pojesti*, however, shows a slightly higher preference for *-do*, while *prepovedati* shows a notably higher preference for *-do*.

5. Genre Comparison

To be able to draw more general conclusions, this section considers only the verbs with a minimum frequency of 10 in all corpora, excluding the Wiki talk subcorpus due to its small size[10]. The following verbs meet this criterion: *biti*, *iti*, *dati*, *vedeti*, *izvedeti*, *zvedeti*, *jesti*, *pojesti*, *povedati*, *odpovedati*, and *prepovedati*.

Having compared all of the genres with one another, we have identified three recurring patterns. In all (sub)corpora, the verb *biti* is predominantly used with the *-do* ending, while the verb *dati* is practically never used with the *-do* ending anymore. Taking into account all the remaining verbs, the blog subcorpus and the Kres corpus generally display comparable tendencies with eight of them (*iti*, *vedeti*, *izvedeti*, *jesti*, *pojesti*, *povedati*, *odpovedati*, and *prepovedati*), whereby the verbs in question show a similar preference for *-do* in some cases, and equal shares of both endings in others.

The second evident pattern is the similarity of the forum, tweet, and news comment subcorpora, as they contain a similar distribution of the endings in the verbs *iti*, *jesti*, *pojesti*, *povedati*, *odpovedati*, and *prepovedati*. In the case of two verbs (*vedeti* and *izvedeti*), the tweet and forum subcorpora show a similar tendency of an equal distribution for both endings.

6. Conclusion

The paper describes the use of endings *-jo* and *-do* in certain Slovene verbs in third person plural. The corpus analysis shows a strong preference for *-do* both in the Kres corpus (12 out of 17 verbs) and in the Janes corpus (10 out of 17 verbs). However, different verbs display completely different tendencies in the analyzed subcorpora, meaning that a general conclusion concerning the use of endings *-jo* and *-do* cannot be made. The verbs *biti* and *dati* proved to be the only ones with the same pattern – *biti* is almost exclusively realized as *bodo*, and *dati* as *dajo*. Regarding the genres in the Janes corpus (tweets, forum posts, blog entries, news comments and on Wikipedia discussions and user pages), a conclusion can be made that there are evident similarities between the blog subcorpus and the Kres corpus, while the tweet, forum, and news comment subcorpora display fairly comparable tendencies as well. Thus, it is important to emphasize that different genres in CMC may not always have the same linguistic characteristics and should therefore not be understood as a homogenous language variety. For a more precise description of the phenomenon, other derivatives (e.g., *spovedati*, *zapovedati*), which were not extracted in the automatic process, should also be included into discussion. The use of the endings *-jo* and *-do* should also be analyzed from a normative perspective, taking into account the language manuals and dictionaries with their descriptions of the analyzed verbs.

7. Acknowledgements

The work described in this paper was funded by the Slovenian Research Agency within the national basic research project “Resources, Tools and Methods for the Research of Nonstandard Internet Slovene” (J6-6842, 2014–2017). We would also like to thank the reviewers for their useful comments and suggestions.

8. References

Arhar Holdt, Š. (2013). Študentje, škratje in nadškofje: končnica *-je* v imenovalniku množine pri samostalnikih prve moške sklanjatve. *Slovenščina 2.0*, 1(1), pp. 134–154.

Crystal, D. (2006). *Language and the Internet*. Cambridge: Cambridge University Press.

Fišer, D., Erjavec, T., Ljubešić, N. (2016). Janes v0.4: korpus slovenskih spletnih uporabniških vsebin. *Slovenščina 2.0* (to appear).

Herring, S.C. (2007). A faceted classification scheme for computer-mediated discourse, *Language@Internet*: http://www.languageatinternet.org/articles/2007/761 (accessed: 13 August 2016).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt Š. and Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; FDV.

Može, S. (2013). Raba kratkega nedoločnika: korpusni pristop. *Slovenščina 2.0*, 1 (1), pp. 155–175.

SP 2001 – Slovenski pravopis. Ljubljana: SAZU – ZRC SAZU – Založba ZRC.

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

[10] For absolute and relative frequencies of verb forms, see https://www.dropbox.com/s/ducs6y7ei75vn9p/Abs_rel.xlsx?dl=0.

Appendix 1[11]

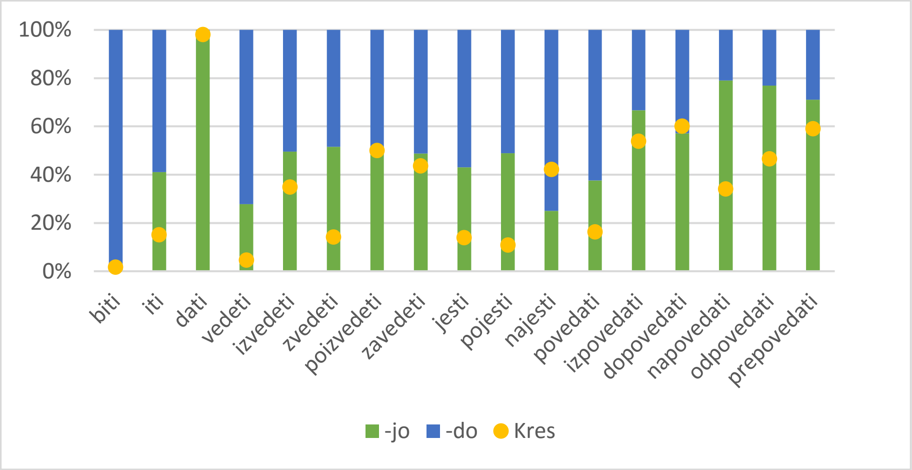

Figure 1: Percentage of *-jo*/*-do* in the tweet subcorpus (green/blue) and the Kres corpus

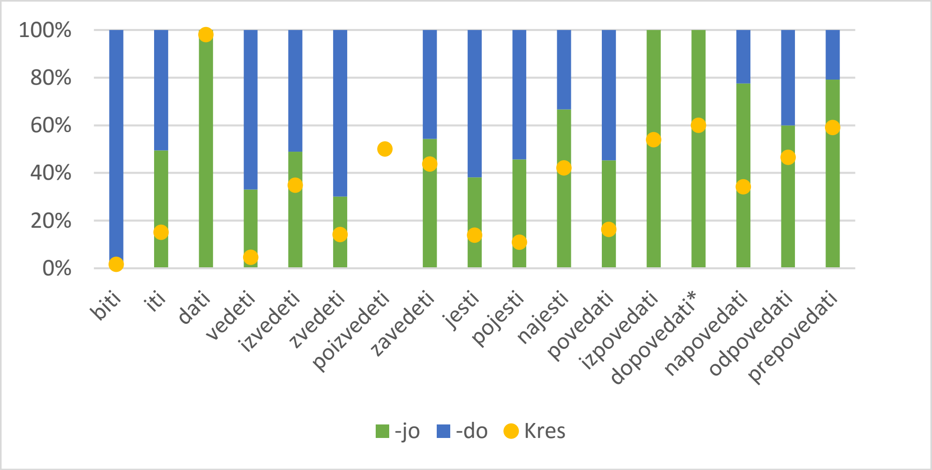

Figure 2: Percentage of *-jo*/*-do* in the forum subcorpus (green/blue) and the Kres corpus (yellow).

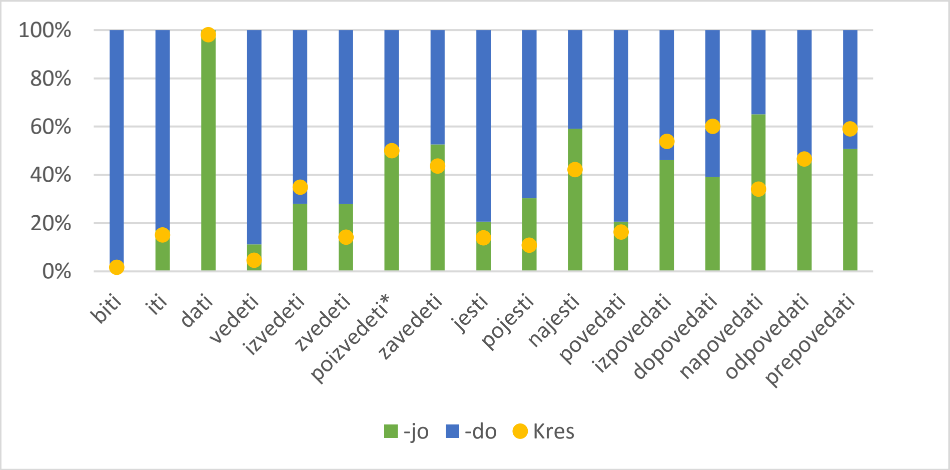

Figure 3: Percentage of *-jo*/*-do* in the blog subcorpus (green/blue) and the Kres corpus (yellow).

[11] The asterisk (*) indicates that one or both verb forms appeared only once. Empty columns indicate that the forms for a particular verb were not found in the Janes corpus. The yellow dot defines the limit between percentage for *-jo* (below it) and *-do* (above it).

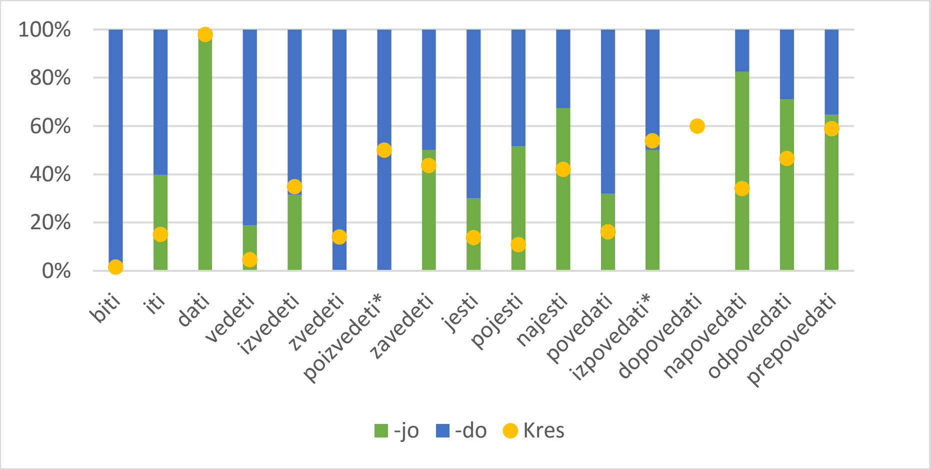

Figure 4: Percentage of *-jo*/*-do* in the news comment subcorpus (green/blue) and the Kres corpus (yellow).

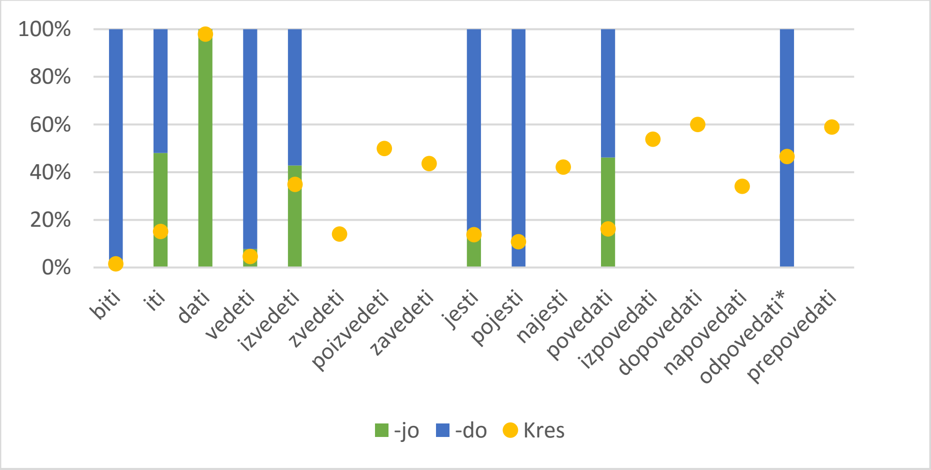

Figure 5: Percentage of *-jo*/*-do* in the Wiki talk subcorpus (green/blue) and the Kres corpus (yellow).