Alternative Endings of Slovene Verbs in Third Person Plural

A Corpus Approach

GAŠPER PESEK

IZA ŠKRJANEC

DAFNE MARKO

Problem description

Slovene → morphologically rich language

Alternative verb endings in 3rd person plural

- *-do*
- *-jo*

5 athematic verbs

- *jejo – jedo* 'they eat'
- *grejo – gredo* 'they go'
- *bojo – bodo* 'they will be'
- *vejo – vedo* 'they know'
- *dajo – dado* 'they give'

Problem description

Slovenski pravopis 2001

- Athematic verbs: *-jo* is deemed less appropriate
- Derivatives of athematic verbs: *-jo* is more commonly used

Alternative endings seem puzzling to Slovene speakers

Analysis of the phenomenon in **CMC** (influenced by spoken language)

Spectrum of genres

Aim of the paper

Determining general tendencies of *-jo* and *-do*

Corpus-based approach

- Janes (UGC)
- Kres (mostly standard written Slovene)

Janes (different genres) vs. Kres

Expected results:

- preference for *-jo* in Janes;
- preference for *-do* in Kres.

Methodology – corpora

2 corpora queried for analysis

Janes v0.4

- UGC
- Over 175 million words
- 9 million documents (2006–2016)
- 5 genres
  - Tweets
  - Forum posts
  - Blog entries + comments
  - Online news + comments
  - Slovene Wikipedia user talk + page talk

Kres

- Standard written Slovene
- Nearly 100 million words
- Over 21,000 documents (1990–2011)
- Balanced genre structure
  - Periodicals (newspapers and magazines)
  - Fiction
  - Non-fiction
  - Documents from the Web
  - Other genres

Methodology – data extraction

Sketch Engine concordancer

1. Verb extraction
   - **CQL:** [word=".+do" & tag="G.*"]
2. Frequency of extracted verbs for *-jo* and *-do* in:
   - the 5 Janes subcorpora;
   - Janes (entire corpus);
   - Kres.
   - **CQL** (e.g., *bojo*)**:** [word="(b|B)ojo" & tag="G.*"]

Methodology – data extraction

17 verbs:

- *biti* ‘be’
- *iti* ‘go’
- *dati* ‘give’
- *vedeti* ‘know’
- *izvedeti* ‘find out’
- *zvedeti* ‘find out’
- *poizvedeti* ‘inquire’
- *zavedeti* ‘realize’
- *jesti* ‘eat’
- *pojesti* ‘eat up’
- *najesti* ‘sate’
- *povedati* ‘tell’
- *izpovedati* ‘confess’
- *dopovedati* ‘get across’
- *napovedati* ‘predict’
- *odpovedati* ‘cancel’
- *prepovedati* ‘forbid’

Analysis – variants in the Janes subcorpora

Athematic verbs

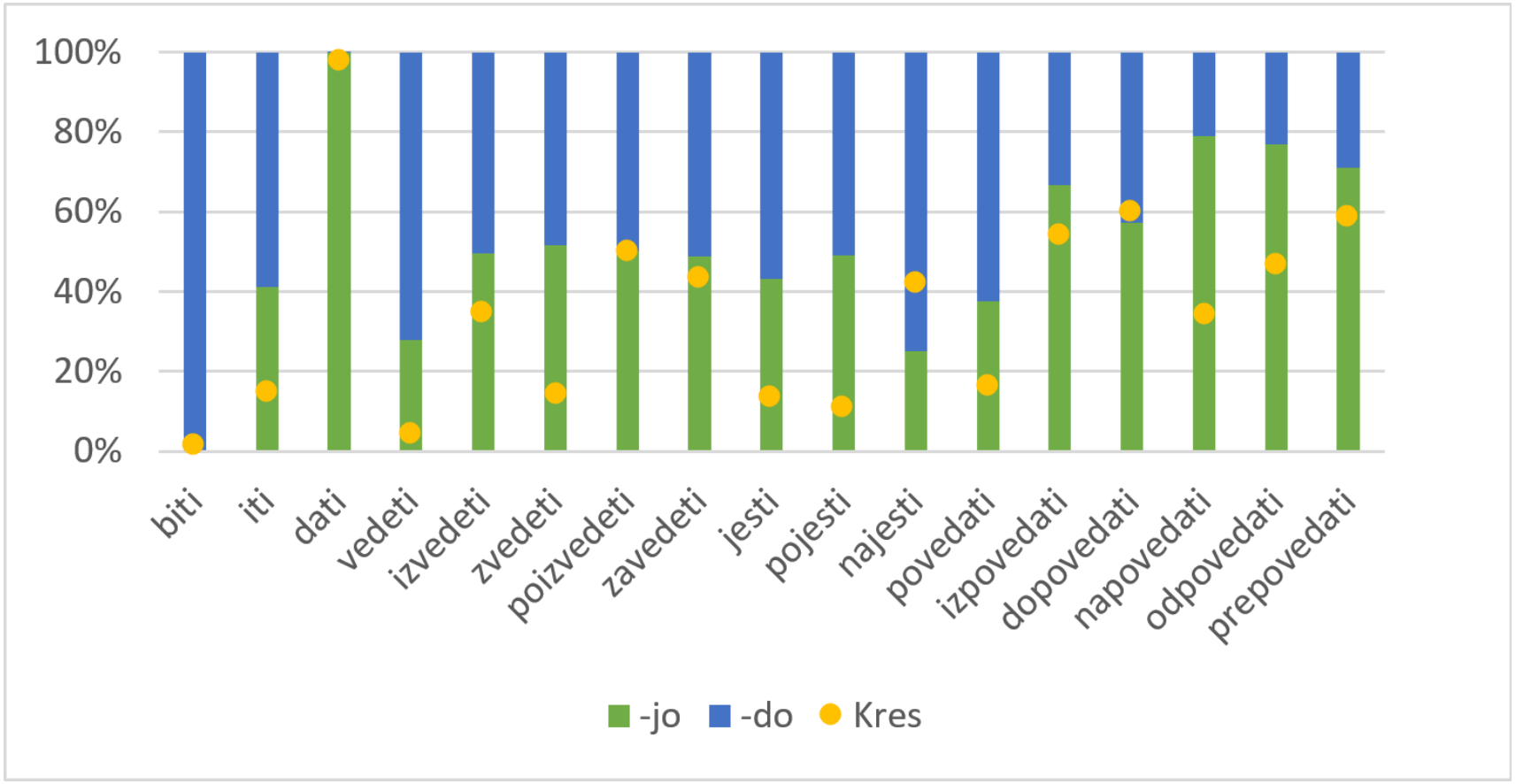
1. *Biti*
   - Virtually 100% of concordances with *-do*
2. *Dati*
   - Virtually 100% of concordances with *-jo*
3. *Vedeti*
   - All subcorpora prefer *-do*
4. *Iti*
   - Equal distribution of *-jo* & *-do*
   - BUT *-do* is preferred in **blogs** (roughly 80%)
5. *Jesti* – general preference for -do
   - Least pronounced in the **tweet** subcorpus
   - Slight preference (roughly 70%) in the **forum** subcorpus
   - Strong preference in the **comment** (roughly 70%), **blog** (80%), and **Wiki talk** (85%) subcorpora

*The following analyses of each individual subcorpus do **not** mention the 5 athematic verbs.*
If general observations are made, however, they are taken into account.

Tweet subcorpus

- *-do* is preferred (60% or more)
  - *najesti*
  - *povedati*
- *-jo* is preferred
  - *izpovedati*
  - *napovedati*
  - *odpovedati*
  - *prepovedati*
- **All other verbs** display a relatively equal distribution of both endings

Tweet subcorpus

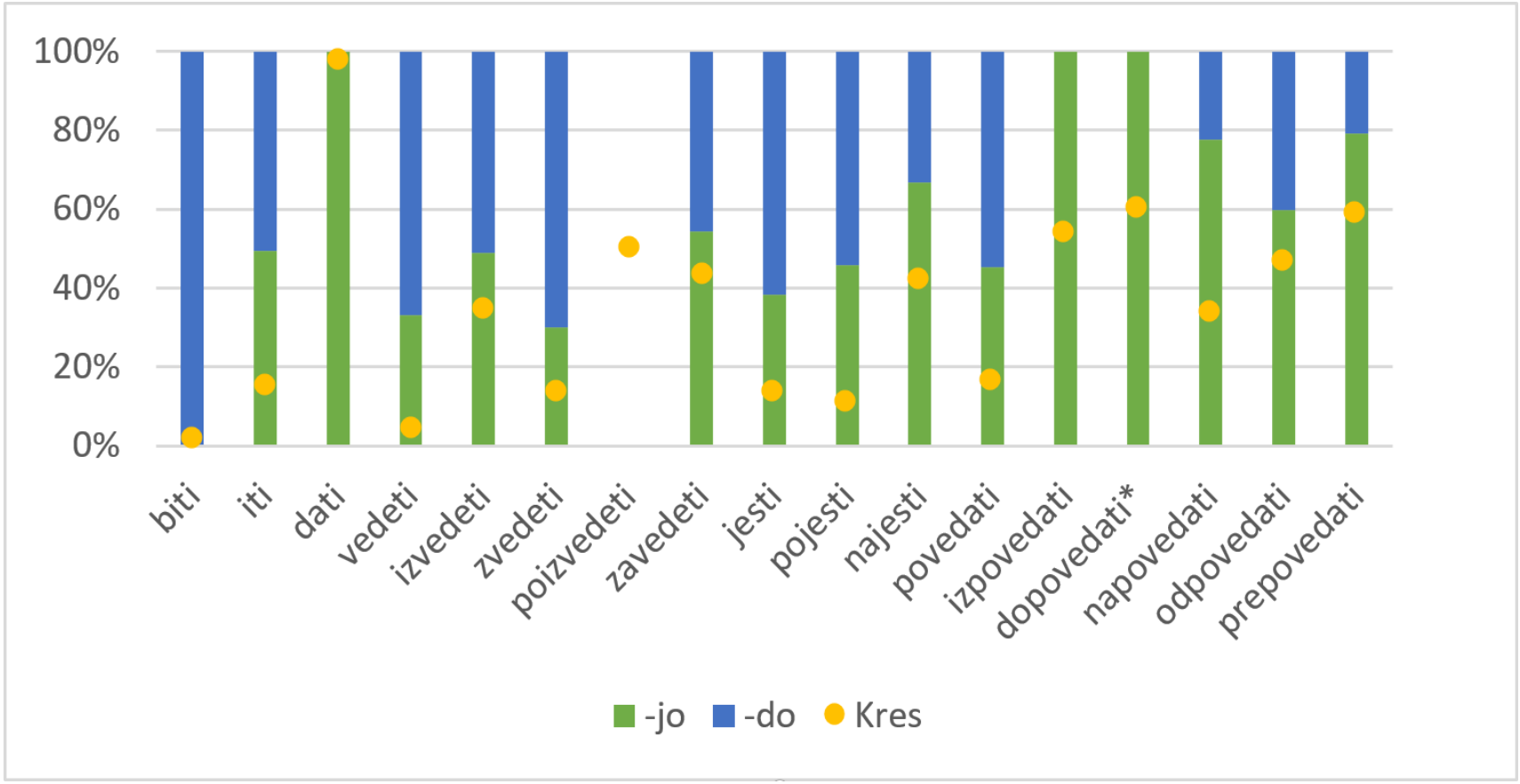


Forum subcorpus

- Considerable preference (roughly 70%) for *-do*
  - *zvedeti*
- A relatively equal distribution of both endings
  - *izvedeti*
  - *zavedeti*
  - *pojesti*
  - *povedati*
- **All other verbs** show a preference (60% or more) for *-jo*
  - *dopovedati** and *izpovedati* have only produced concordances with *-jo*

* one or both verb forms appeared only once

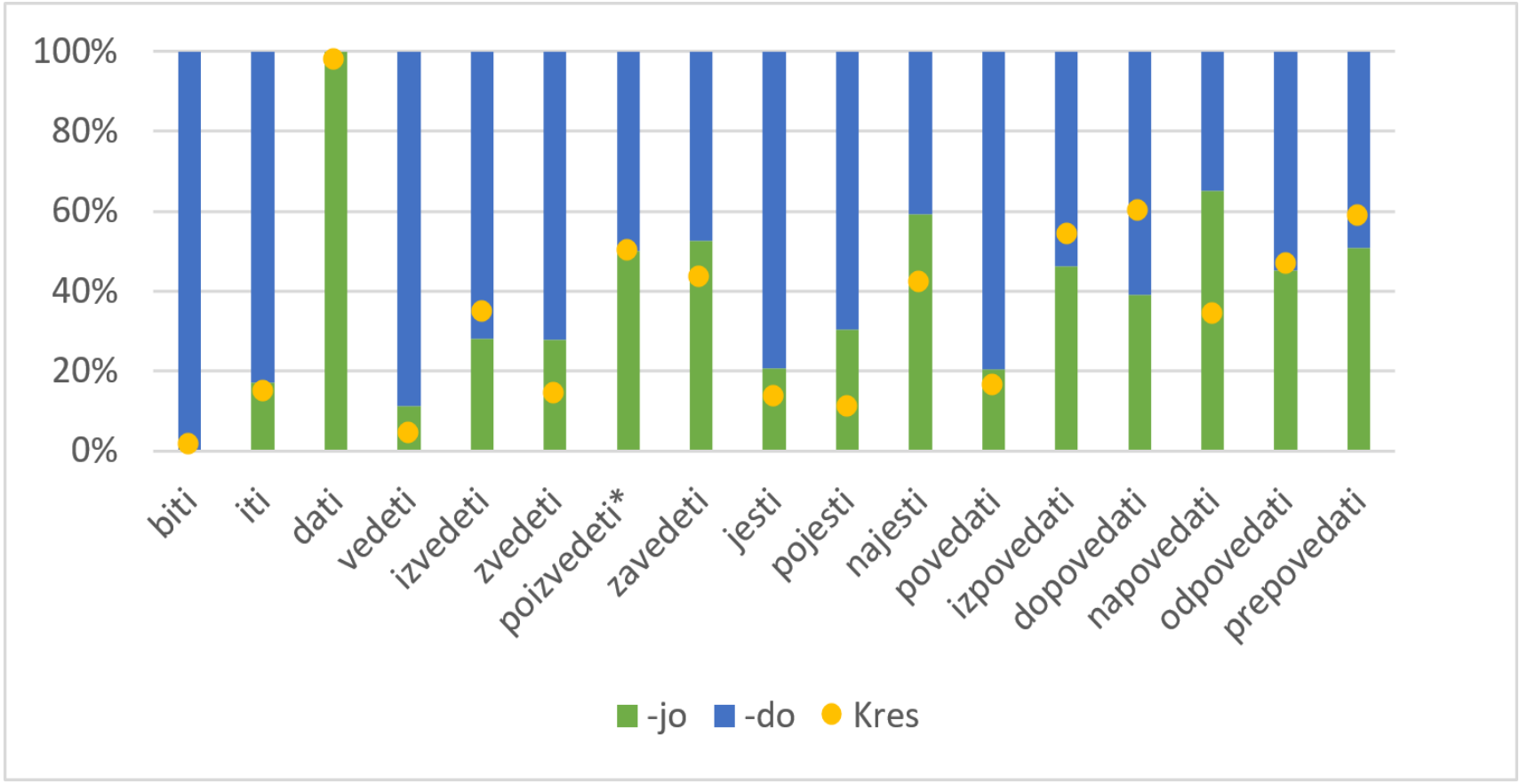
Forum subcorpus



Blog subcorpus

- Overall preference for *-do* with these (non-athematic) verbs:
  - *izvedeti*
  - *zvedeti*
  - *pojesti*
  - *povedati*
  - *dopovedati*
- A relatively even distribution of both endings
  - *poizvedeti**
  - *zavedeti*
  - *najesti*
  - *izpovedati*
  - *odpovedati*
  - *prepovedati*
- A notable preference (over 60%) for *-jo*
  - *napovedati*

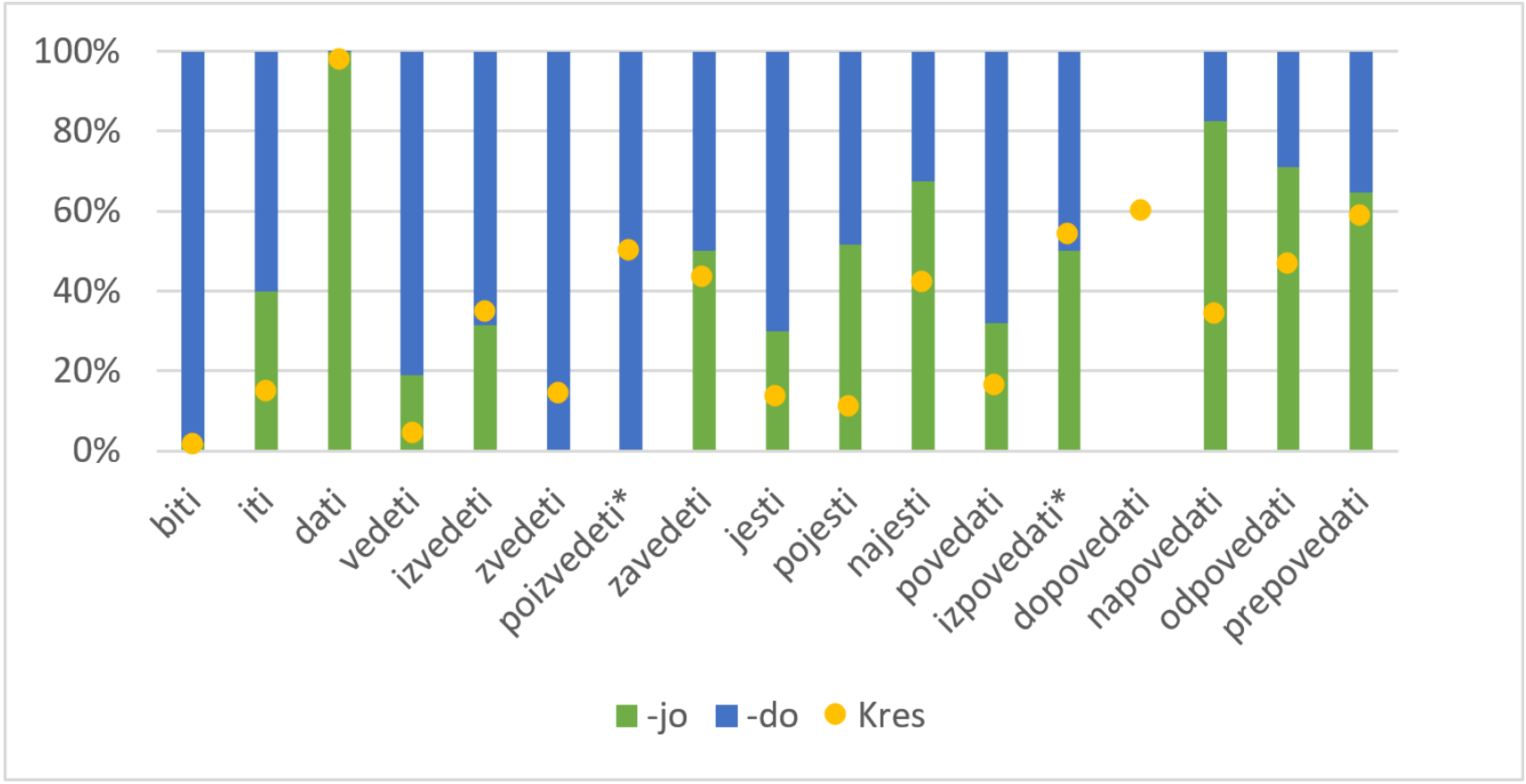
Blog subcorpus



News comment subcorpus

- A notable preference (60% or more) for *-do*
  - *izvedeti*
  - *povedati*
  - *zvedeti* and *poizvedeti** have only provided concordances with *-do*
- A relatively equal distribution of the two endings
  - *zavedeti*
  - *pojesti*
  - *Izpovedati**
- A notable preference for *-jo*
  - *najesti*
  - *napovedati*
  - *odpovedati*
  - *prepovedati*

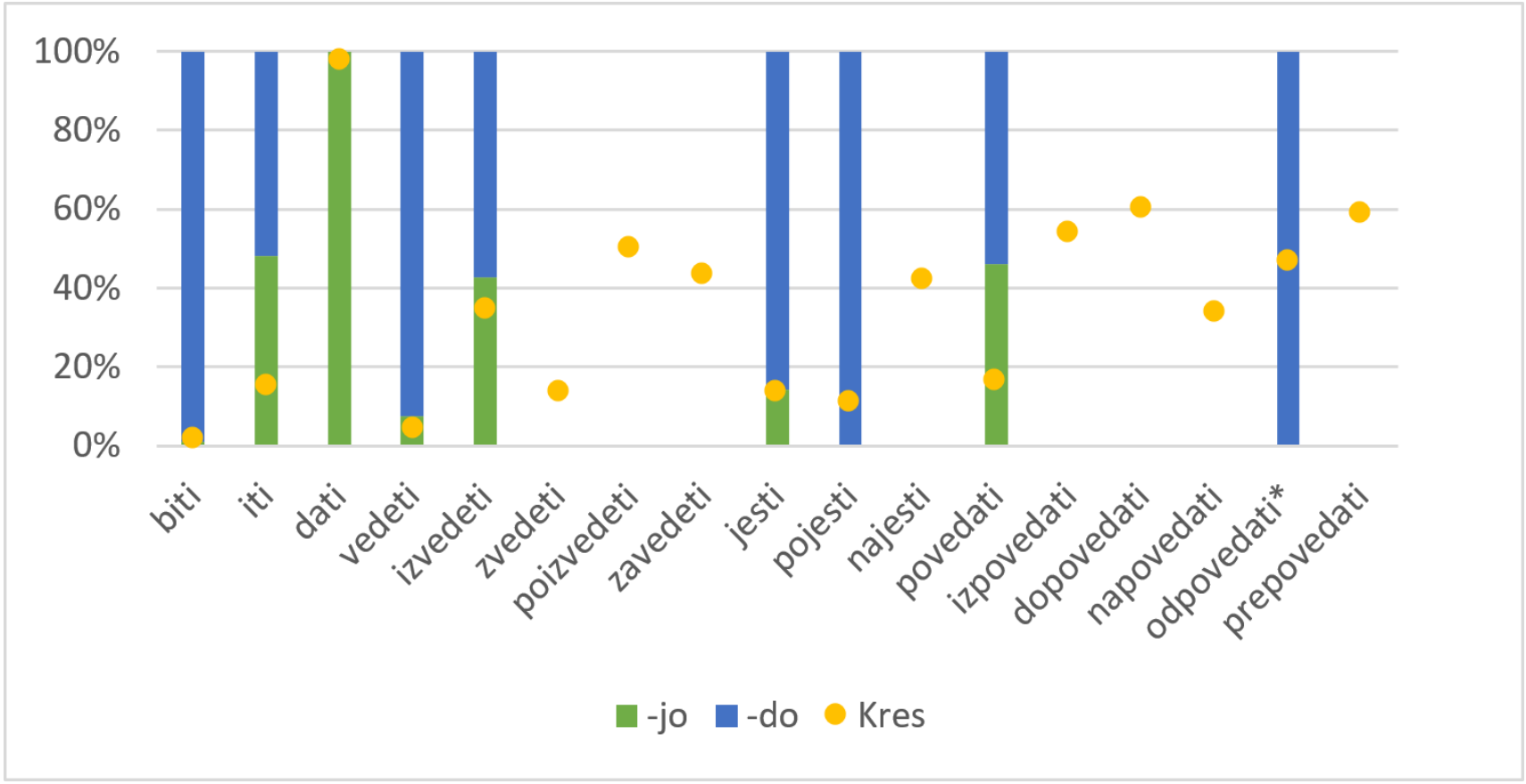
News comment subcorpus



Wiki talk subcorpus

- Few verbs produced any concordances at all
- Verbs with concordances show a strong overall preference for *-do*
- A relatively even distribution of the two endings
  - *povedati*
  - *izvedeti*

Wiki talk subcorpus



Analysis – the Janes subcorpora vs. Kres

Athematic verbs

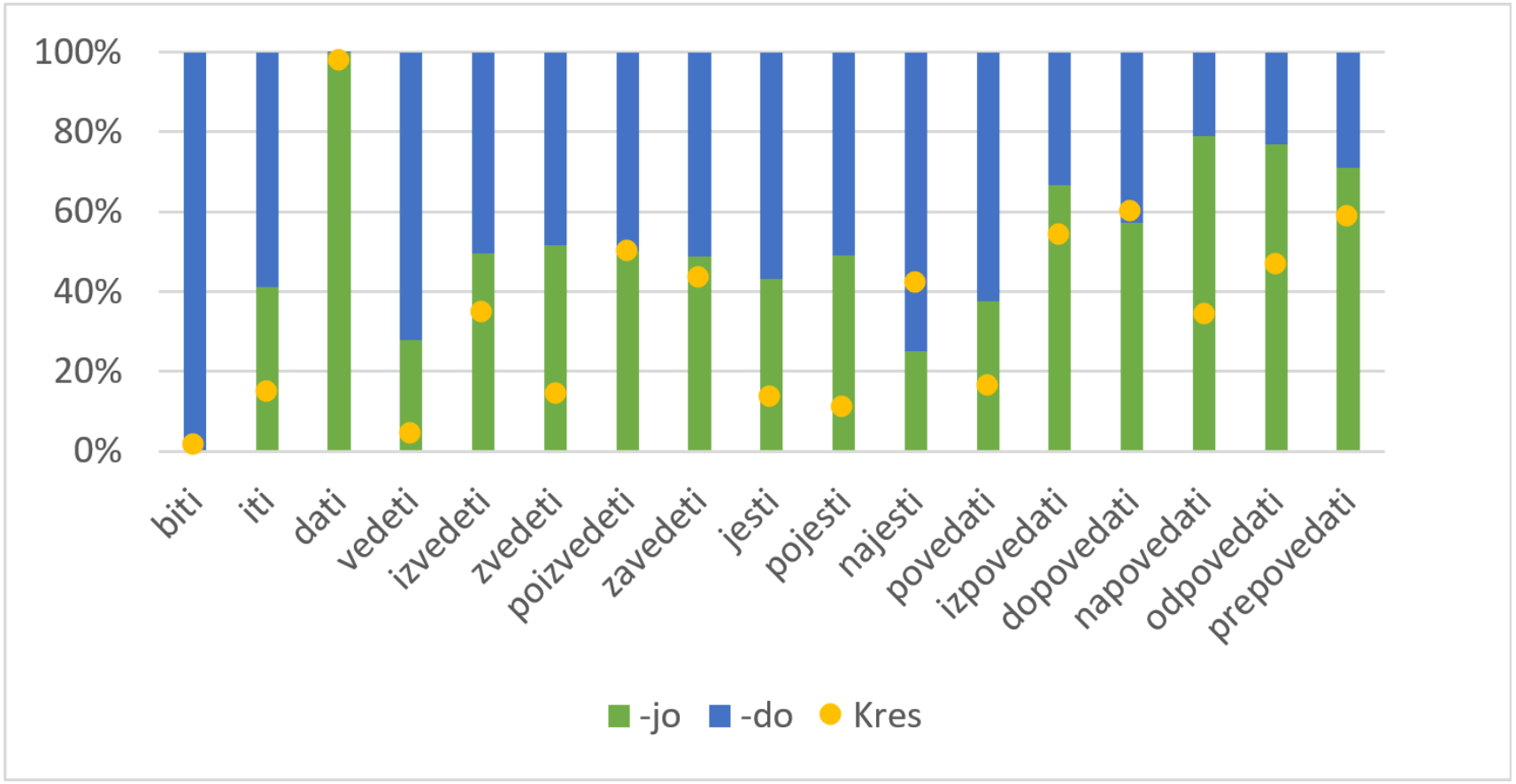
1. *Biti*
   - The *-jo*/*-do* ratios overlap
2. *Dati*
   - *-jo* is almost exclusively employed in both corpora
3. *Vedeti*
   - The **tweet**, **forum**, and **news comment** subcorpora prefer *-jo* compared to Kres
   - The **blog** and **Wiki talk** subcorpora are similar to Kres (shows a slightly stronger preference for *-do*)
4. *Iti*
   - Most Janes subcorpora → a much stronger preference for *-jo*
   - BUT *-jo*/*-do* ratios are comparable in the **blog** subcorpus
5. *Jesti*
   - A notable preference for *-jo* in the **tweet**, **forum**, and **news comments** subcorpora
   - An overlap with Kres ratio in the **Wiki talk** subcorpus
   - The **blog** subcorpus slightly prefers -jo compared to Kres

*The following analyses of each individual subcorpus do **not** mention the 5 athematic verbs.*
If general observations are made, however, they are taken into account.

Tweet subcorpus vs. Kres

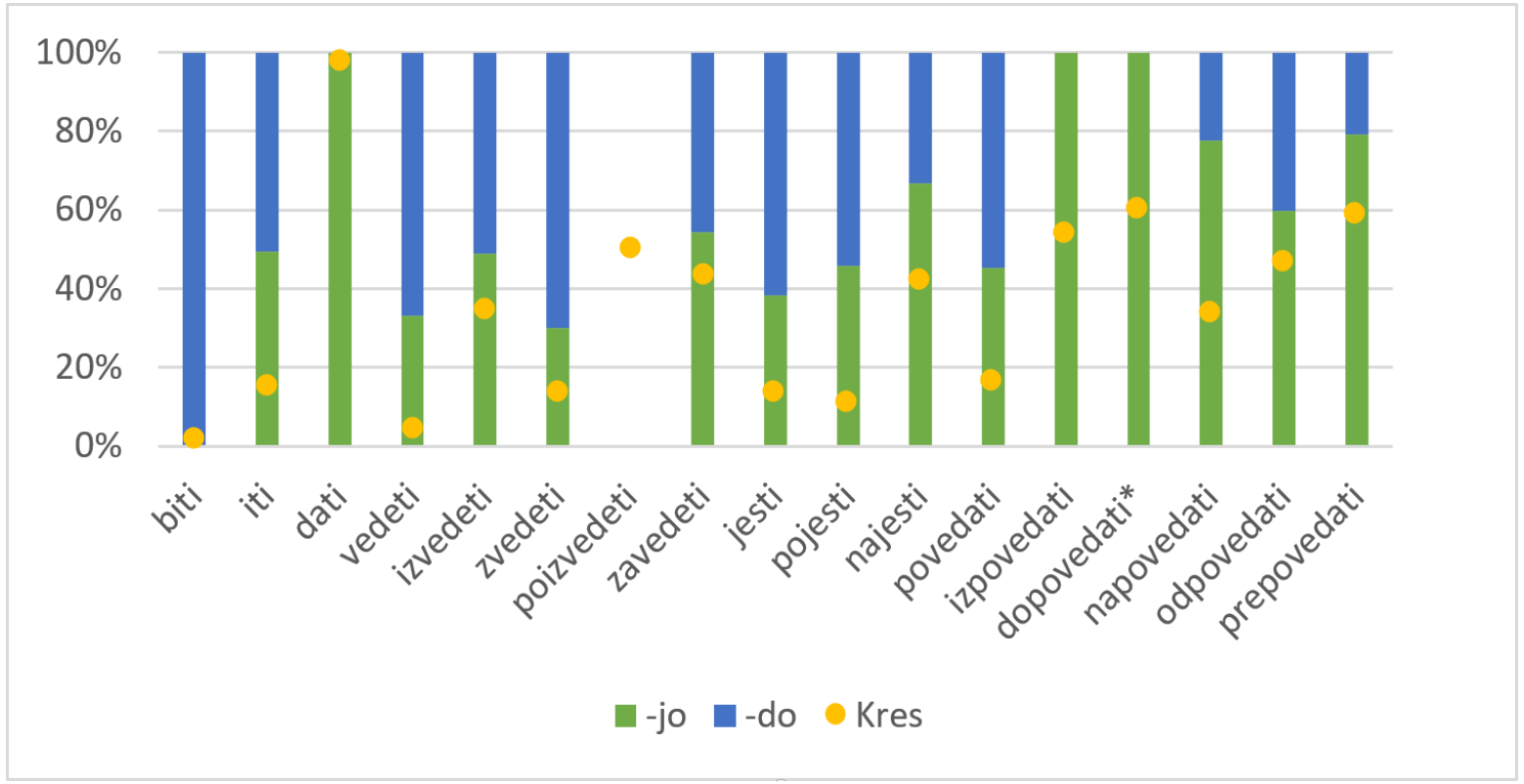
- Similar ratios
  - *poizvedeti*
  - *zavedeti*
  - *dopovedati*
- With *najesti*, there is a higher preference for *-do* in the **tweets**
- **All other verbs** show a notable to strong preference for *-jo* in comparison with Kres

Tweet subcorpus vs. Kres



Forum subcorpus vs. Kres

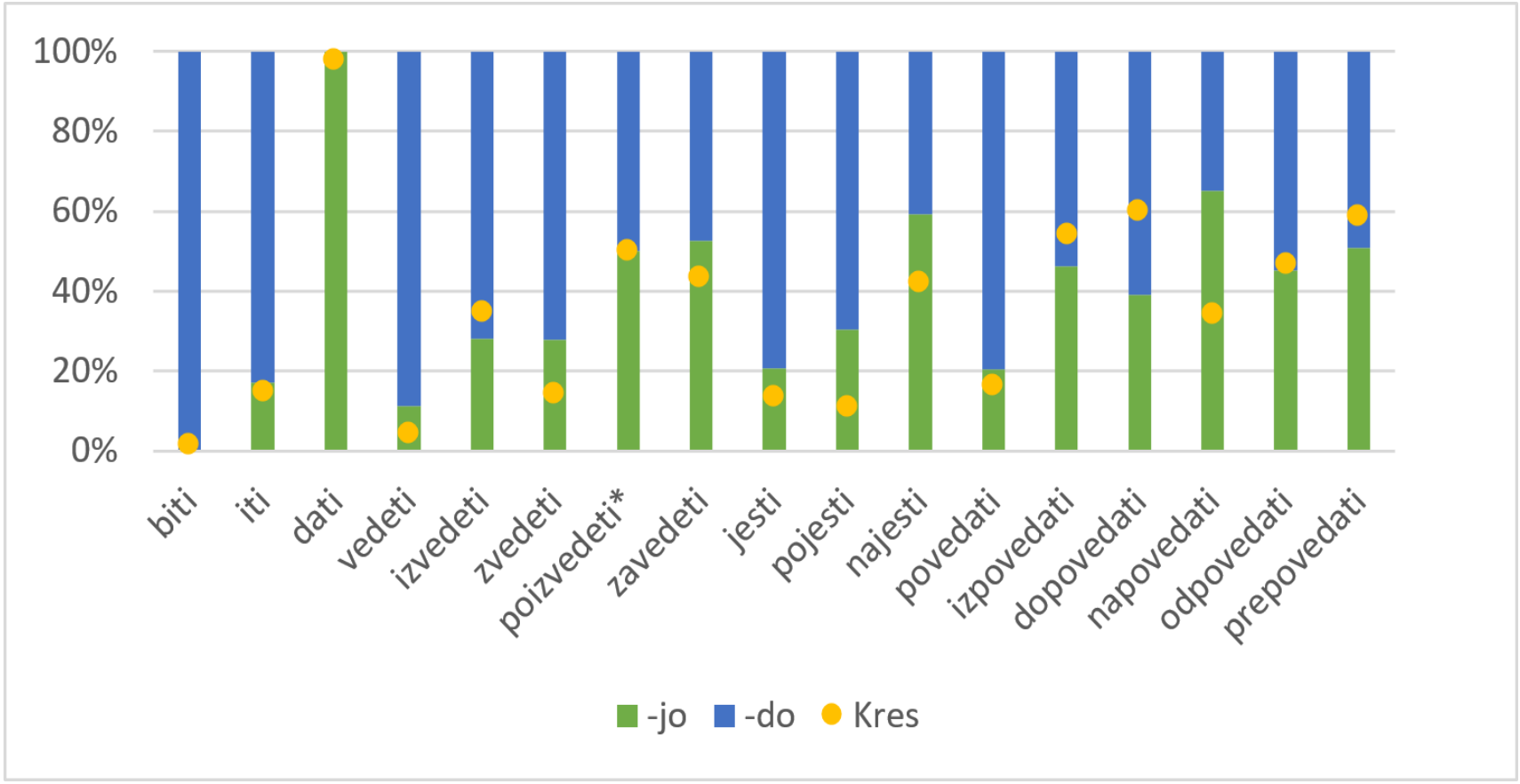
- All verbs show a higher preference for *-jo* (in varying degrees)



Blog subcorpus vs. Kres

- A slightly higher preference for *-jo* with the verb *zavedeti*
- A stronger preference for *-jo*
  - *zvedeti*
  - *pojesti*
  - *najesti*
  - *napovedati*
- Similar ratios
  - *poizvedeti**
  - *povedati*
  - *odpovedati*
- The **remaining verbs** (*izvedeti*, *izpovedati*, *dopovedati*, and *prepovedati*) seem to prefer *-do* more than Kres.

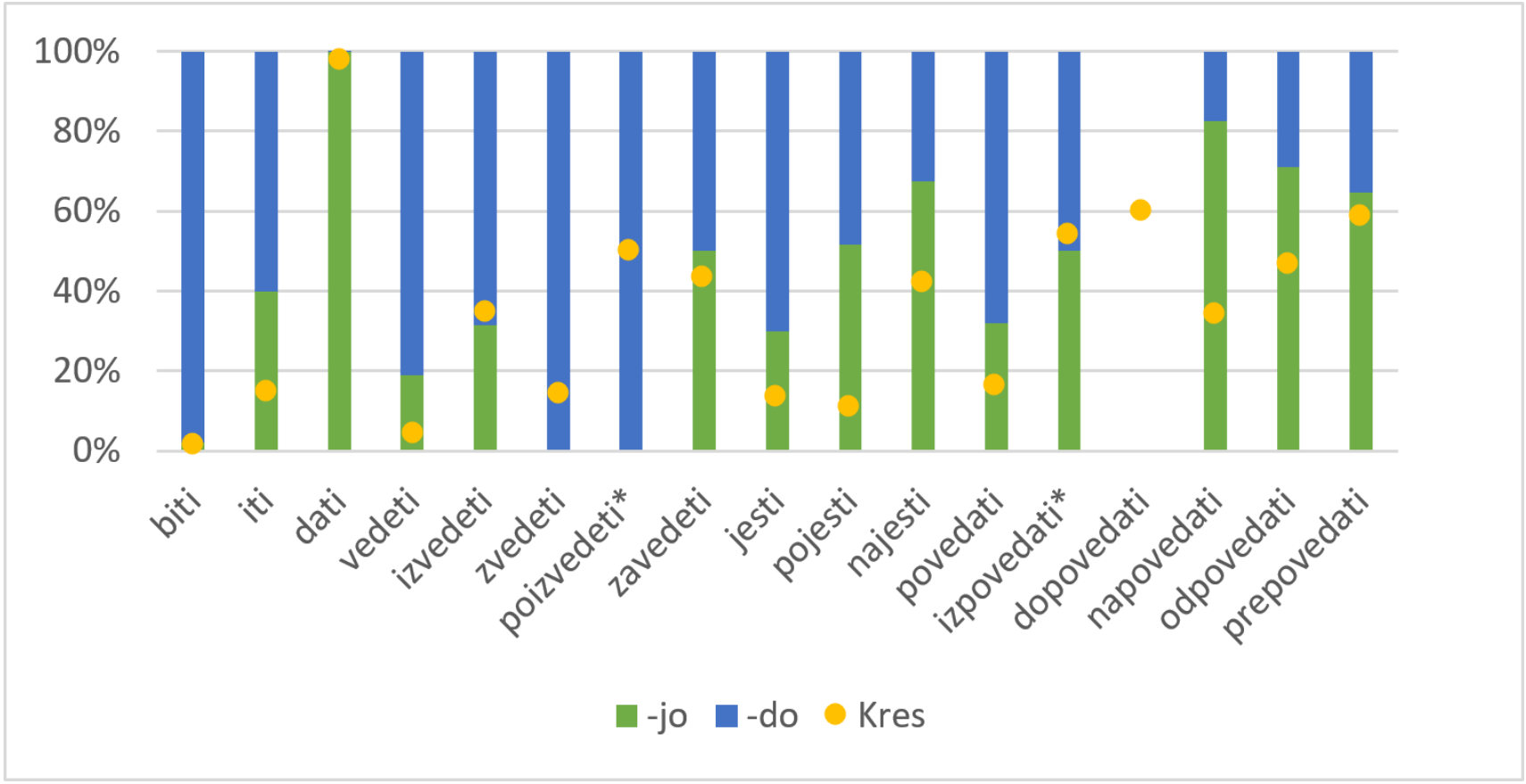
Blog subcorpus vs. Kres



News comment subcorpus vs. Kres

- A higher preference for *-jo*
  - *pojesti*, *najesti*, *povedati*, *napovedati*, and *odpovedati*
- Similar ratios
  - *izvedeti*, *zavedeti*, *izpovedati**, and *prepovedati*
- A higher preference for *-do*
  - *zvedeti* and *poizvedeti**

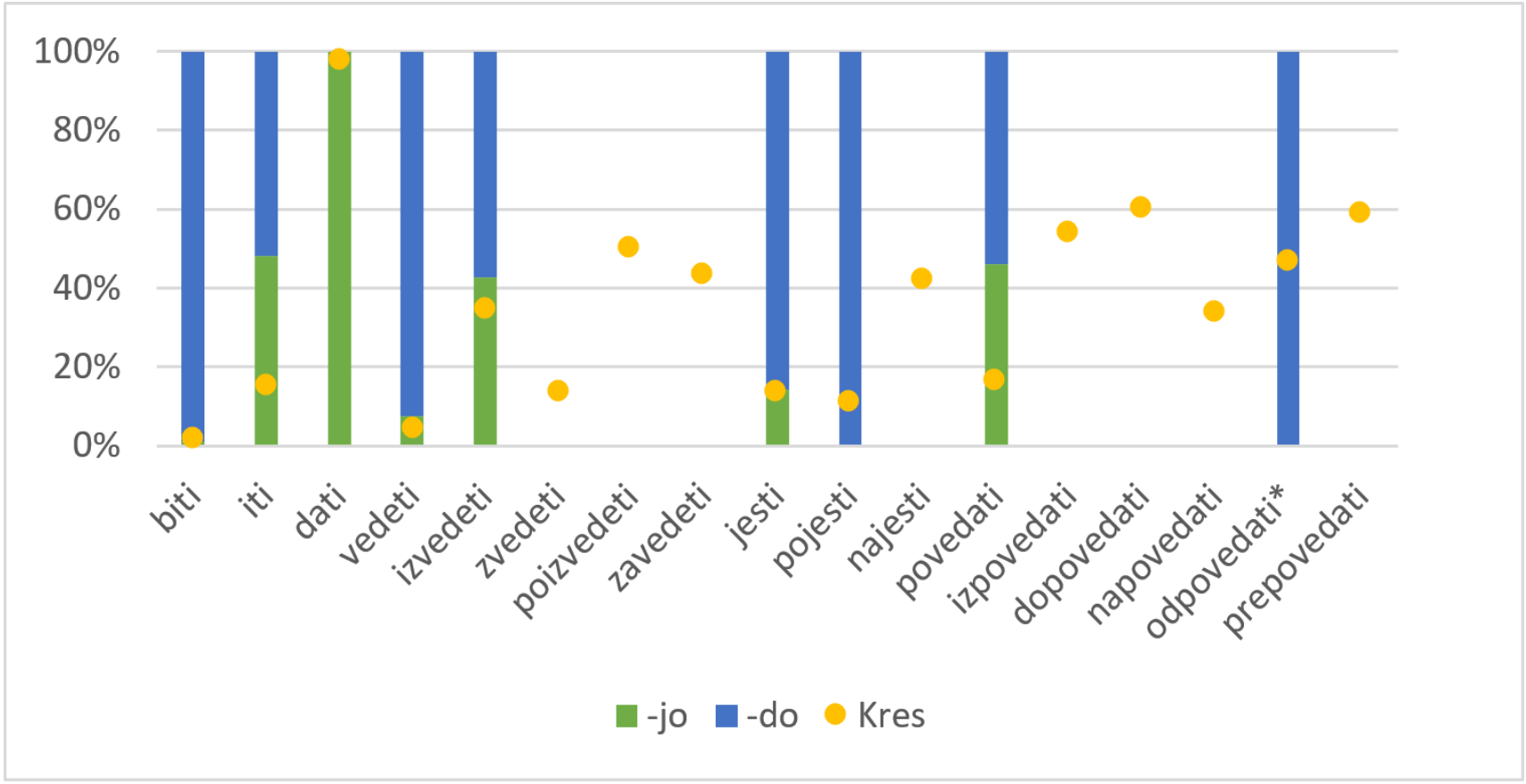
News comment subcorpus vs. Kres



Wiki talk subcorpus vs. Kres

- A slightly higher preference for *-jo* with *izvedeti*
- A notably higher preference for *-jo* with *povedati*
- A slightly higher preference for *-do* with *pojesti*
- A notably higher preference for *-do* with *prepovedati*

Wiki talk subcorpus vs. Kres



Genre comparison

- Wiki talk subcorpus excluded (too small)*
- Minimum frequency of 10 in all corpora

- **Verbs that met this criterion:**
  - *biti*
  - *iti*
  - *dati*
  - *vedeti*
  - *izvedeti*
  - *zvedeti*
  - *jesti*
  - *pojesti*
  - *povedati*
  - *odpovedati*
  - *prepovedati*

* absolute and relative frequencies for all verb forms in all (sub)corpora: https://www.dropbox.com/s/ducs6y7ei75vn9p/Abs_rel.xlsx?dl=0

Genre comparison

- Recurring patterns:
  - *biti* almost exclusively employs *-do* in all (sub)corpora;
  - *dati* almost exclusively employs *-jo* in all (sub)corpora;
  - the **blog** subcorpus and the **Kres** corpus display comparable tendencies with 8 verbs;
  - similarity of the **forum**, **tweet**, and **news comment** subcorpora.

Conclusion

- A preference for -*do* has been found in both corpora
  - Janes (10 out of 17 verbs)
  - Kres (12 out of 17 verbs)
- Different verbs show different tendencies → generalizations mostly not possible
- *Biti* almost universally realized as *bodo*
- *Dati* almost universally realized as *dajo*
- Genre comparison:
  - evident similarities between the **blog subcorpus** and **Kres**;
  - the **tweet**, **forum**, and **news comment** subcorpora also show similarities.

Further work

- Analyzing other derivatives (e.g., *spovedati*, *zapovedati*)
- Analyzing *-jo* and *-do* from a normative perspective

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.

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.