cjvt Center za jezikovne vire in tehnologije

How Users Responded to a Responsive Dictionary: The Case of the Thesaurus of Modern Slovene

Špela Arhar Holdt, Jaka Čibej

Centre for Language Resources and Technologies, University of Ljubljana

(Faculty of Arts, Faculty of Computer and Information Science)

Zagreb, 10. 5. 2019

Univerza v Ljubljani Fakulteta za računalništvo in informatiko

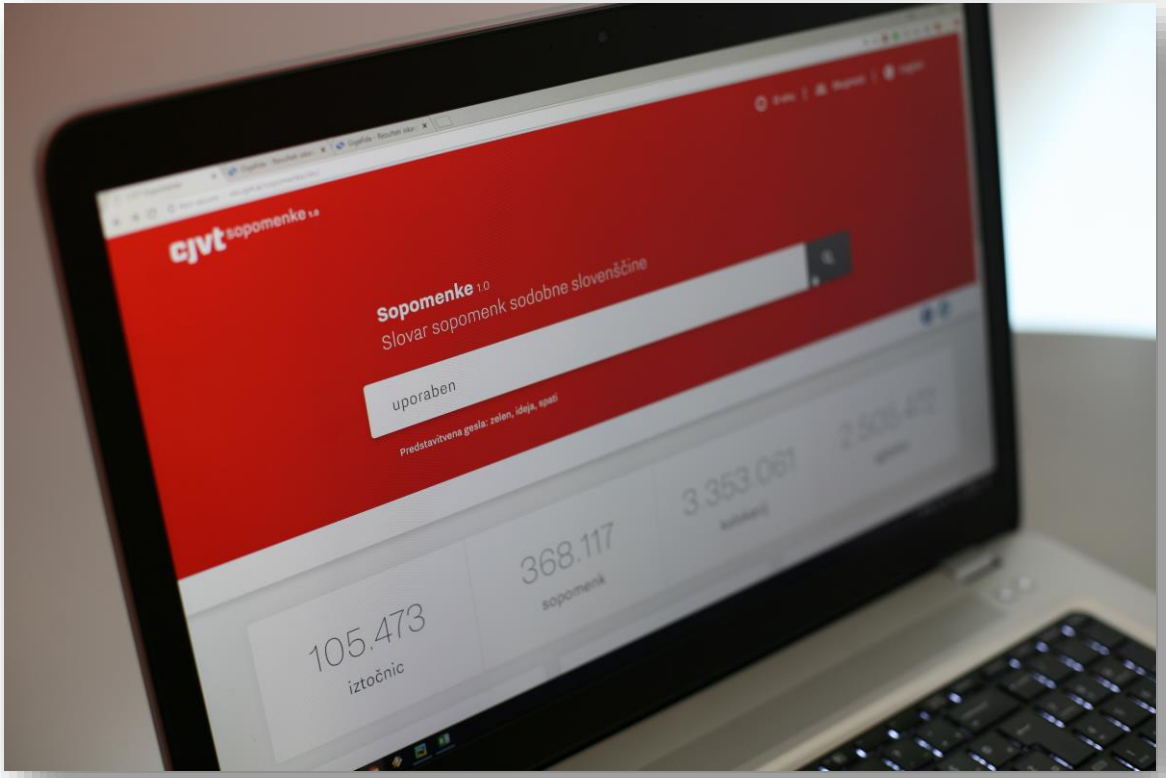
Univerza v Ljubljani Filozofska fakulteta

REPUBLIKA SLOVENIJA MINISTRY OF CULTURE

THE THESAURUS OF MODERN SLOVENE

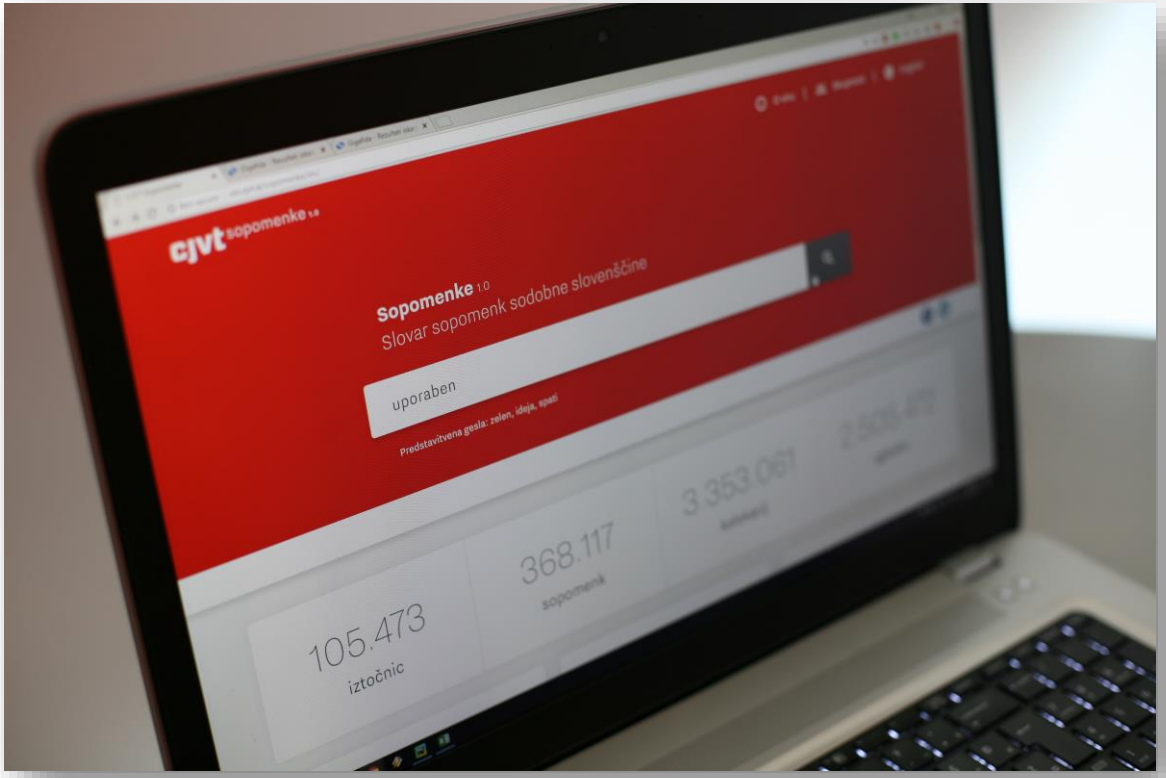
The Thesaurus of Modern Slovene

- Published in March 2018: http://viri.cjvt.si/sopomenke/eng/
- 105,473 headwords, 368,117 synonyms, 3,353,061 collocations and 2,505,472 corpus examples = largest **open-access** collection of Slovene synonyms
- Fully **automatically generated** from existing language resources
- **Users can help** clean out the noise in the data and contribute new synonyms to the resource



The Thesaurus of Modern Slovene

- Short **tutorials** and **recorded lectures** on the CJVT Youtube channel (with English subtitles).
- Promotional material: https://www.cjvt.si/promocija-sopomenk/
- Concise information inside the interface.
- The methodology of the database compilation:
  Krek et al. 2017b
- Interface design, user involvement, user tracking:
  Arhar Holdt et al. 2018



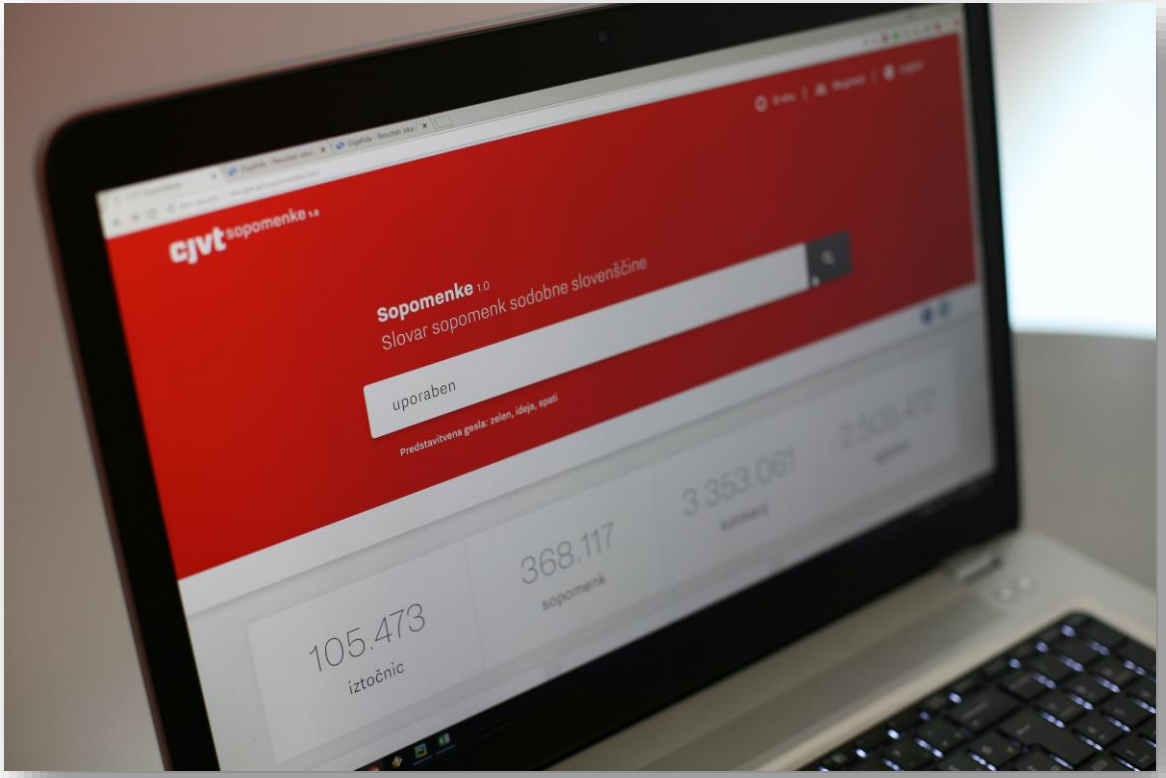
Responsive Dictionary

- Developed specifically for the **digital medium**.
- The database is constructed (semi-)**automatically.**
- and released to the community as soon as it is linguistically evaluated as relevant (albeit somewhat noisy).
- **The community can contribute** towards improving the resource.
- A clearly defined methodology for including user contributions and information collected from user activities in the dictionary.
- **The data evolves,** the changes are transparently tracked and archived.
- The database is **openly accessible** under an appropriate license.

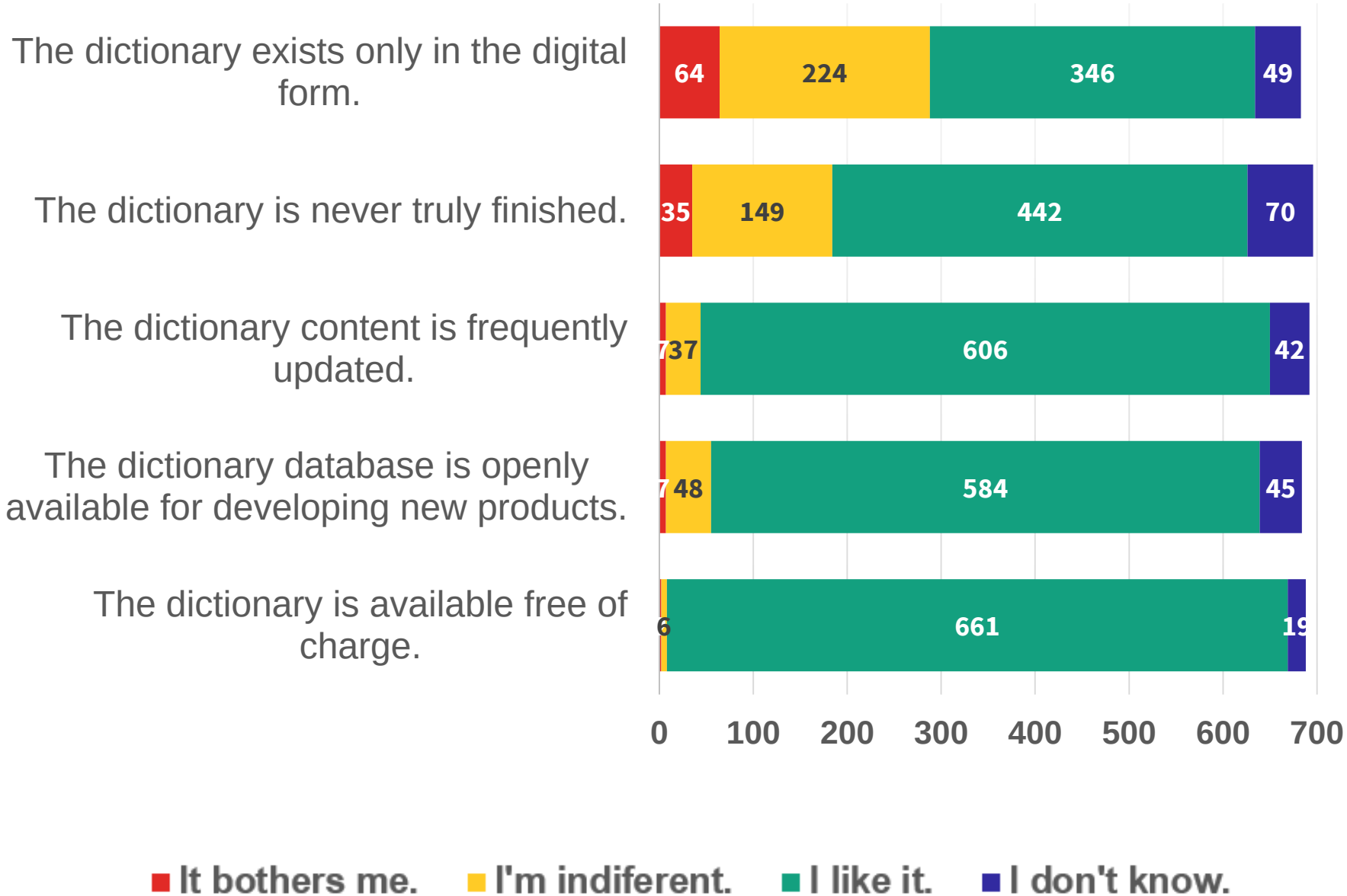
= Responds to language change and the feedback from the language community.

User Survey

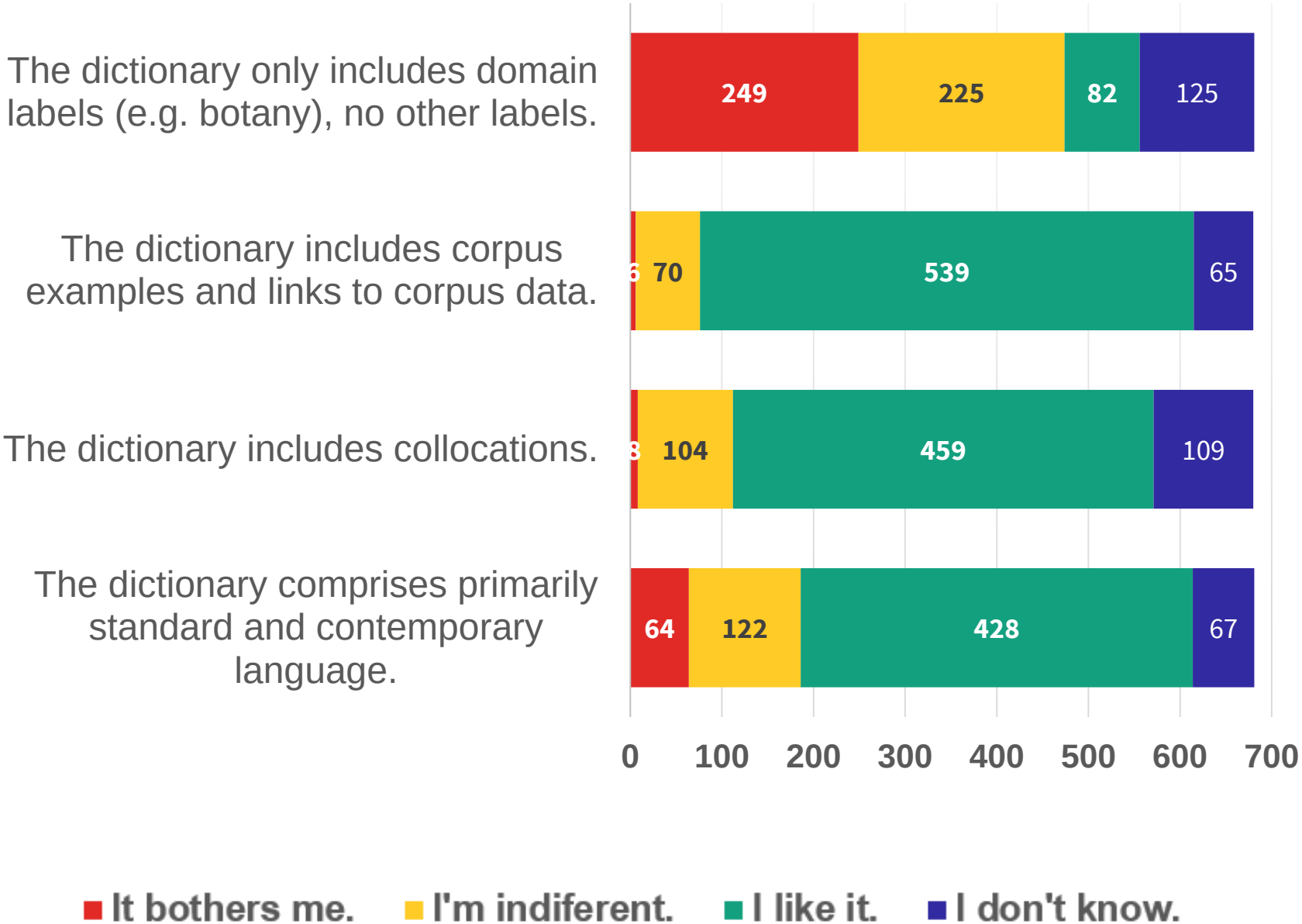
- How does the language community **percieve these novelties?**
- A survey funded by the Slovene Ministry of Culture.
- Online questionnaire, digital format (https://www.1ka.si/)
- 7 questions (cca. 3 minutes)
- **671 completed**, 956 partially completed
- Numerus comparable to the study by Müller Spitzer et al. (Using Online Dictionaries, 2014)

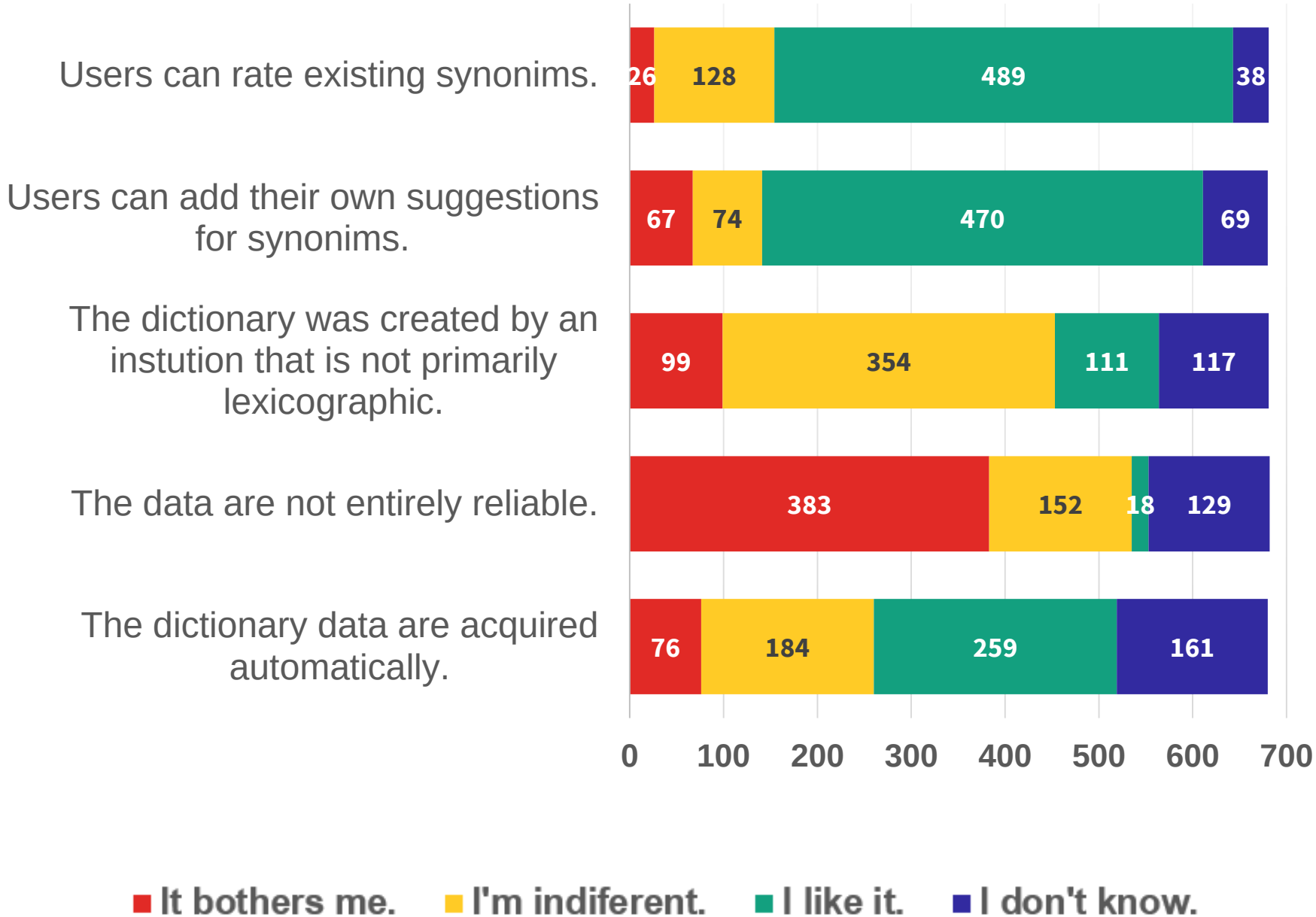


WHAT IS YOUR OPINION ON THE FOLLOWING (14) FEATURES?


The dictionary exists only in the digital form.
64
224
346
49
The dictionary is never truly finished.
35
149
442
70
The dictionary content is frequently updated.
7
37
606
42
The dictionary database is openly available for developing new products.
7
48
584
45
The dictionary is available free of charge.
6
661
19
0
100
200
300
400
500
600
700
It bothers me.
I'm indiferent.
I like it.
I don't know.

cjvt


The dictionary only includes domain labels (e.g. botany), no other labels.
249
225
82
125
The dictionary includes corpus examples and links to corpus data.
6
70
539
65
The dictionary includes collocations.
8
104
459
109
The dictionary comprises primarily standard and contemporary language.
64
122
428
67
0
100
200
300
400
500
600
700
It bothers me.
I'm indiferent.
I like it.
I don't know.


Users can rate existing synonims.
26
128
489
38
Users can add their own suggestions for synonims.
67
74
470
69
The dictionary was created by an instution that is not primarily lexicographic.
99
354
111
117
The data are not entirely reliable.
383
152
18
129
The dictionary data are acquired automatically.
76
184
259
161
0
100
200
300
400
500
600
700
It bothers me.
I'm indiferent.
I like it.
I don't know.

STATISTICAL ANALYSIS

Methodology

- We prepared contingency tables including **participants‘ opinions** and their **metadata** (see below).
- We ran a **chi square** test of independence, then calculated **Pearson residuals** to see if there are any statistically significant differences between the groups.
- For this presentation: **Familiarity with the Thesaurus** and **Age.**

- **[Status** (36% employed in the public sector, 6% in the private sector, 8% secondary school students, 7% university students etc.) and **Interest in language** (352 participants are professionally involved in language editing, 332 in translating, 120 teach Slovene in primary schools, 75 in secondary schools, etc.)] - User typology defined by Arhar Holdt et al. 2016.

Contingency Tables: Example

„The dictionary comprises primarily **standard and contemporary language“** vs. Familiarity with the Thesaurus.

[1] "CONTINGENCY TABLE"

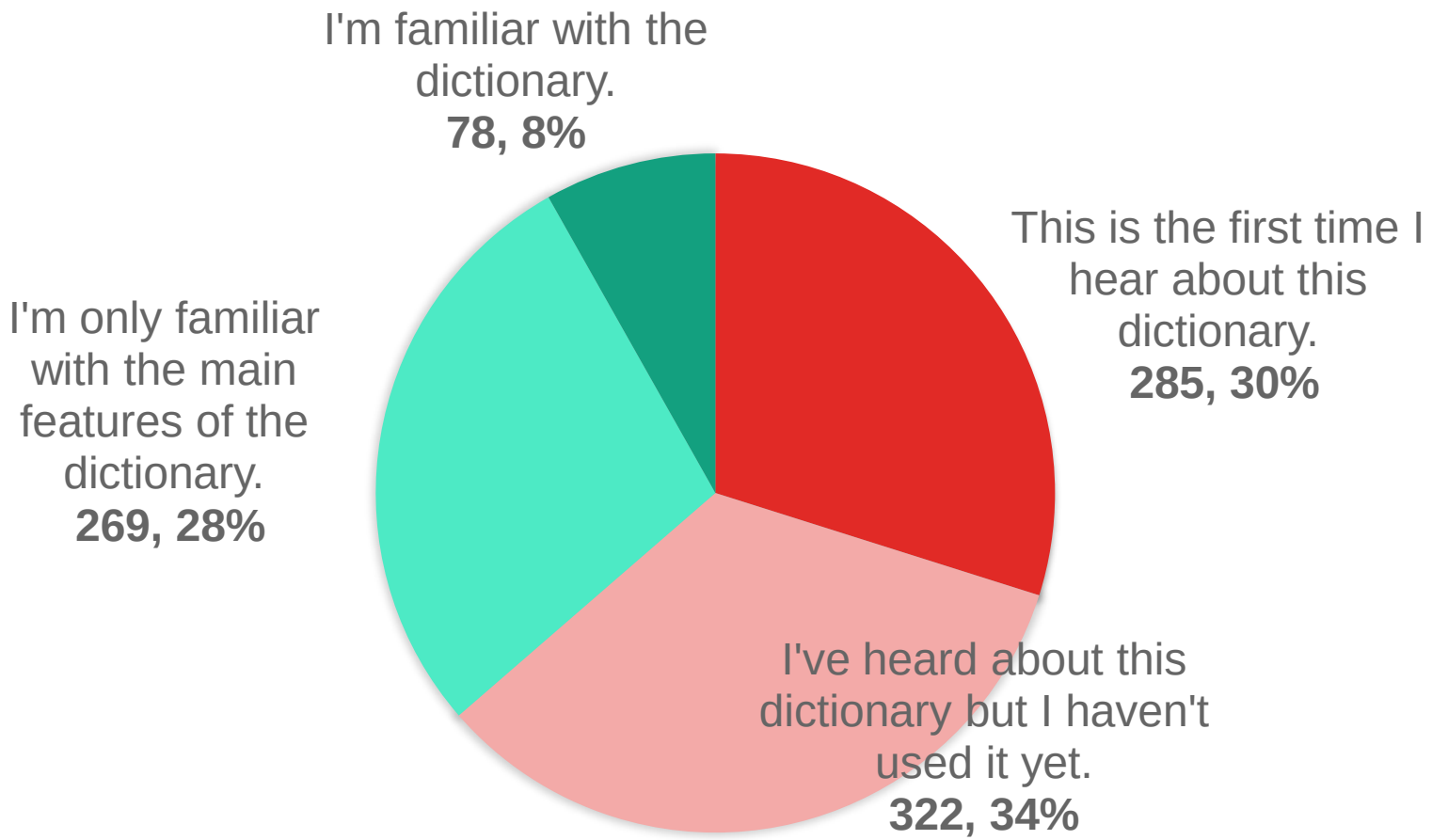
	I'm indifferent.	I don't know.	I like it.	It bothers me.
I don't know the Thesaurus.	67	41	252	26
I know the Thesaurus.	54	25	176	38

[1] "PEARSON RESIDUALS"

	I'm indifferent.	I don't know.	I like it.	It bothers me.
I don't know the Thesaurus.	-0.215	0.568	0.557	-1.721
I know the Thesaurus.	0.247	-0.652	-0.639	1.976

[1] "χ2 = 8.441, N = 679, df = 3, p = 0.038"
[1] "Difference significant: * - Cramer's V: 0.11"
[1] "Significance info: **** - p <= 0.0001; *** - p <= 0.001; ** - p <= 0.01; * - p <= 0.05"

HOW FAMILIAR ARE YOU WITH THE THESAURUS OF MODERN SLOVENE?



Familiarity with the Thesaurus vs. Opinions (1)

Standard vocabulary [$\chi^2 = 8.441$, N = 679, df = 3, p = 0.038]
The participants who are familiar with the Thesaurus are slightly more frequently bothered by the fact that the Thesaurus only includes modern and standard vocabulary ($r_P = 1.976$).

Collocations [$\chi^2 = 44.071$, N = 678, df = 3, p ~ 0]
The participants who are unfamiliar with the Thesaurus rarely like the fact that the Thesaurus includes collocation data ($r_P = -2.393$), and are either indifferent ($r_P = 2.028$) or undecided ($r_P = 3.023$) about it. On the other hand, the participants who are familiar are rarely indifferent ($r_P = -2.325$) or undecided ($r_P = -3.465$), and are more in favor of collocations ($r_P = 2.744$).

Corpus examples [$\chi^2 = 26.569$, N = 678, df = 3, p ~ 0]
The participants who are unfamiliar with the Thesaurus are more frequently undecided ($r_P = 2.910$) about whether the inclusion of corpus examples in the Thesaurus is good or not. The opposite is true of the participants who are familiar with it ($r_P = -3.345$).

Familiarity with the Thesaurus vs. Opinions (2)

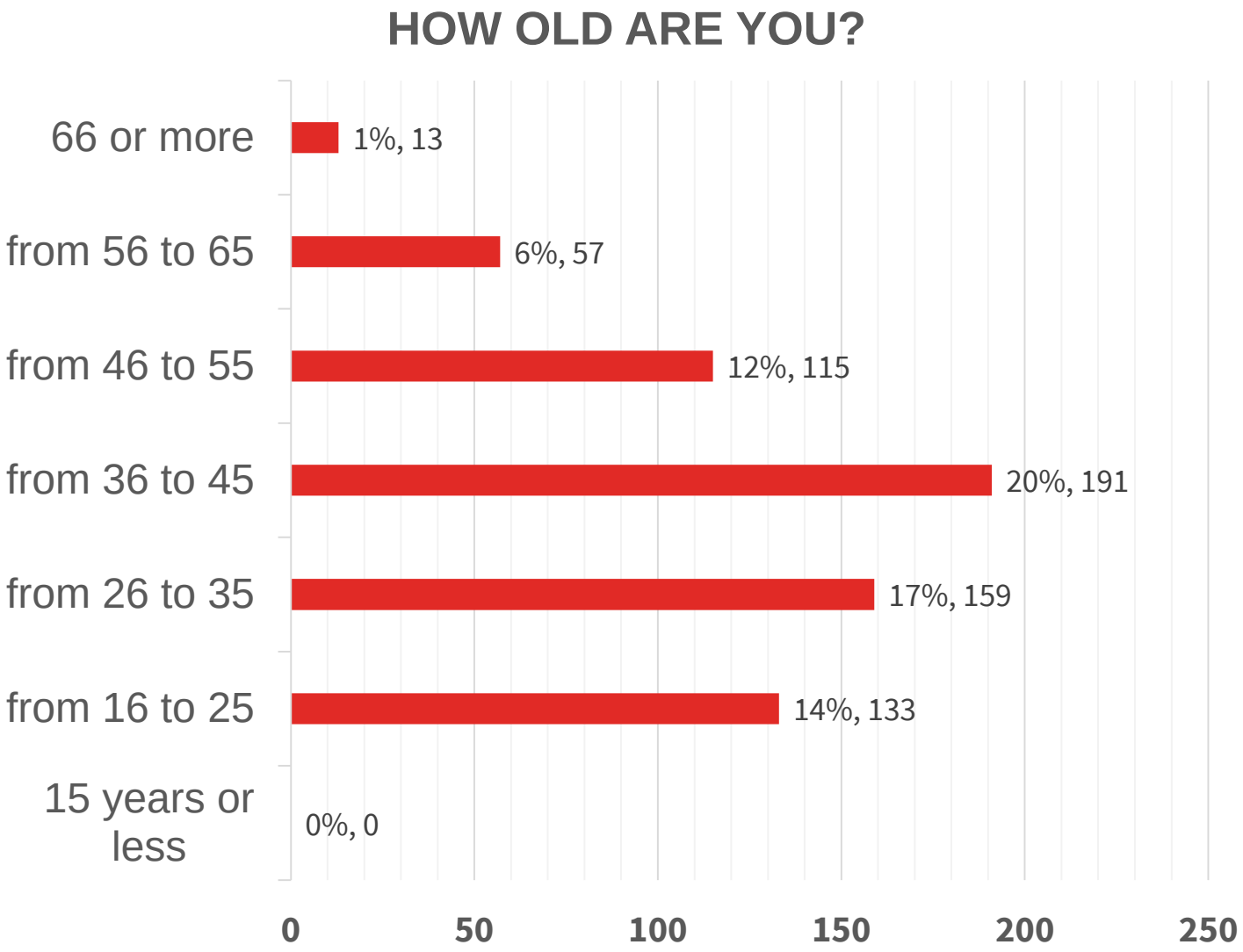
User votes [$\chi^2 = 9.407$, N = 679, df = 3, p = 0.024]
The majority of the participants like user votes. The participants who are familiar with the Thesaurus are rarely undecided whether user votes are good or not ($r_P = -2.244$).

User suggestions [$\chi^2 = 17.38$, N = 678, df = 3, p = 0.001]
Similiarly, the participants who are familiar with the Thesaurus are rarely undecided whether user suggestions are good or not ($r_P = -2.285$).

Labels [$\chi^2 = 19.972$, N = 679, df = 3, p ~ 0]
The participants who are unfamiliar with the Thesaurus are less frequently bothered ($r_P = -2.148$) by the fact that the Thesaurus contains no labels apart from domain labels. The participants who are familiar with the Thesaurus, on the other hand, are rarely undecided about it ($r_P = -2.120$) and more frequently bothered by it ($r_P = 2.456$).

Other opinions (e.g. the automaticity of the approach, the reliability of the data, the institution behind the dictionary) show no correlations: the familiarity with the dictionary does not play a role.

cjvt


HOW OLD ARE YOU?
66 or more
1%, 13
from 56 to 65
6%, 57
from 46 to 55
12%, 115
from 36 to 45
20%, 191
from 26 to 35
17%, 159
from 16 to 25
14%, 133
15 years or less
0%, 0
0
50
100
150
200
250

Age vs. Opinions (1)

Never finished [$\chi^2 = 39.283$, N = 660, df = 15, p = 0.001]
Compared to other age groups, the participants between the ages of 16 and 25 are either indifferent ($r_P = 3.714$) to the fact that the Thesaurus is never truly finished or are bothered by it ($r_P = 2.214$). They rarely like it ($r_P = -2.928$).

User suggestions [$\chi^2 = 28.07$, N = 656, df = 15, p = 0.021]
The 16-25 age group is either indifferent ($r_P = 2.581$) to user suggestions in the Thesaurus or bothered by them ($r_P = 2.582$).

Frequently updated [$\chi^2 = 27.832$, N = 662, df = 15, p = 0.023]
The 16-25 age group is more frequently indifferent ($r_P = 3.876$) to the fact that the Thesaurus is frequently updated.

Institution [$\chi^2 = 32.35$, N = 657, df = 15, p = 0.006]
The 16-25 age group likes ($r_P = 2.596$) the fact that the Thesaurus was made by an institution that is not primarily lexicographic.

Age vs. Opinions (2)

Labels [$\chi^2 = 41.497$, N = 657, df = 15, p ~ 0]
The 16-25 age group likes ($r_P = 3.503$) the fact that the Thesaurus contains no labels apart from domain labels and is rarely bothered by it ($r_P = -3.164$). The 36-45 age group rarely likes the lack of labels ($r_P = -2.020$).

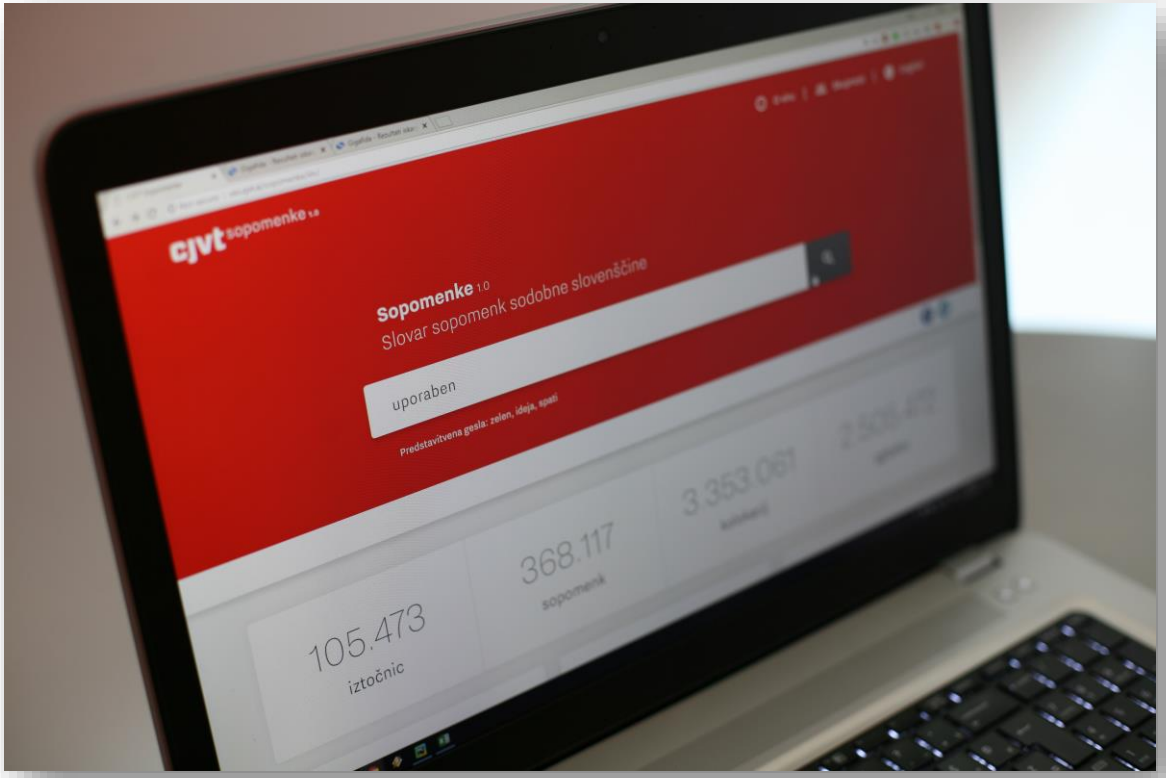
Corpus examples [$\chi^2 = 27.844$, N = 656, df = 15, p = 0.023]
The 16-25 age group is more frequently indifferent ($r_P = 3.289$) to the fact that the Thesaurus includes corpus examples and links to corpus data. The same is true for the group 66 years and above ($r_P = 2.285$).

Only digital [$\chi^2 = 36.65$, N = 657, df = 15, p = 0.001]
The 46-55 age group is more frequently undecided ($r_P = 2.757$) whether a digital-only dictionary is good or not.

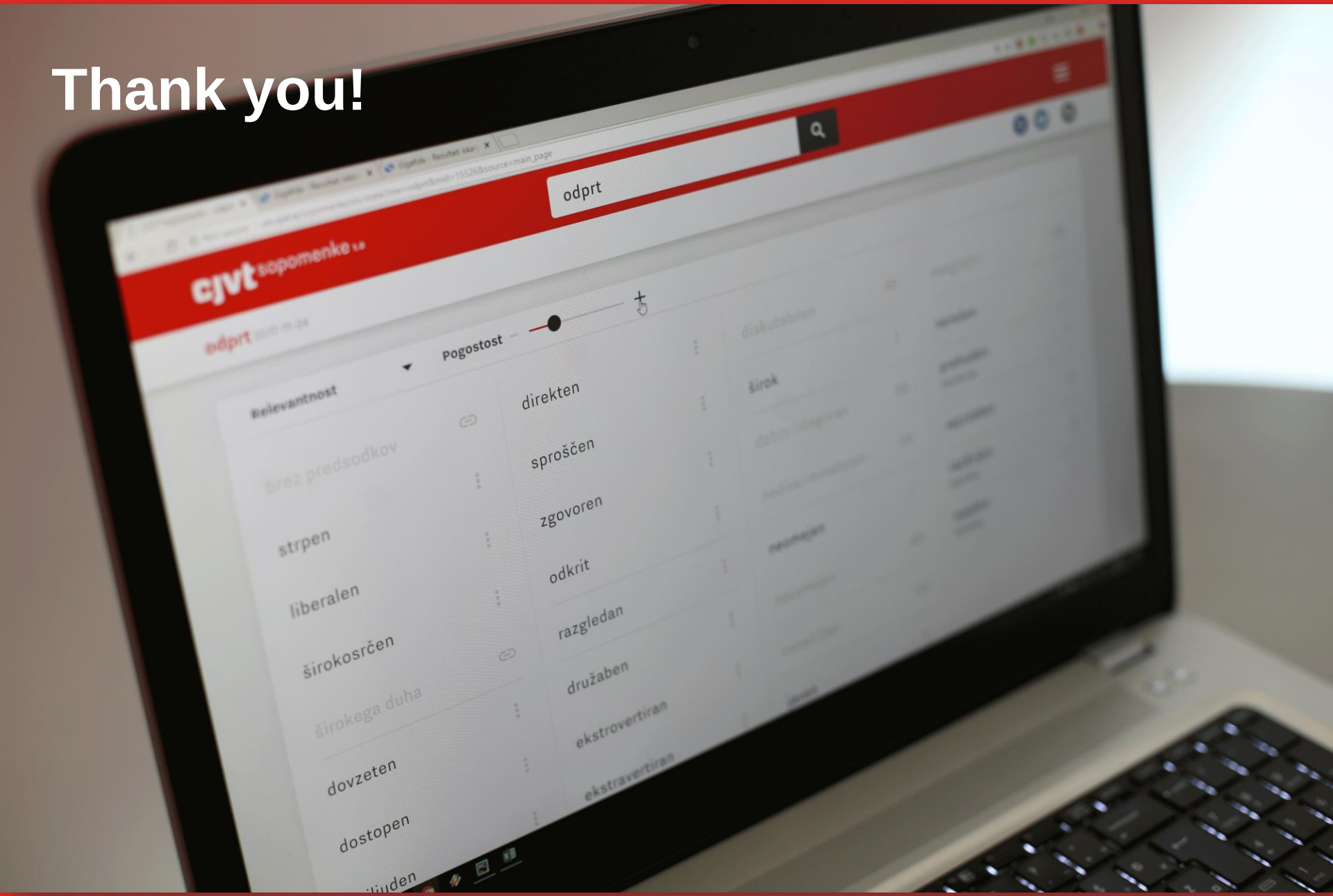
Automatically compiled [$\chi^2 = 31.45$, N = 656, df = 15, p = 0.008]
The 36-45 age group is rarely indifferent ($r_P = -2.310$) to the fact that the Thesaurus has been compiled automatically.

And finally …

- An open-ended section where participants expressed their opinions and suggestions for further improvements.
- 186 comments – constructive and useful suggestions.
- **Users want:** additional data (non-standard vocabulary and labels), interface improvements, connectibility with other resources, etc.
- Survey findings will be implemented in the upcoming updates of the Thesaurus.



Thank you!


odprt
cjvt sopomenke
Relevantnost
Pogostost
direkten
sproščen
zgovoren
odkrit
razgledan
družaben
ekstrovertiran
strpen
liberalen
širokosrčen
dovzeten
dostopen

cjvt

References

- Krek et al. (eds.). 2017a. Sopomenke 1.0: Thesaurus of Modern Slovene. viri.cjvt.si/sopomenke.
- Arhar Holdt, Špela, Jaka Čibej, Kaja Dobrovoljc, Apolonija Gantar, Vojko Gorjanc, Bojan Klemenc, Iztok Kosem, Simon Krek, Cyprian Laskowski, Marko Robnik Šikonja. 2018. Thesaurus of Modern Slovene: By the Community for the Community. In: Čibej, Jaka, Vojko Gorjanc, Iztok Kosem, Simon Krek (eds.). Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. ISBN 978-961-06-0097-8). 1st ed. Ljubljana: ZIFF, pp. 401-410.
- Krek, Simon, Cyprian Laskowski, Marko Robnik Šikonja. 2017b. From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. In: Kosem, Iztok (ed.) et al., Proceedings of eLex 2017: Lexicography from Scratch, 19-21 September 2017, Leiden, Netherlands.
- Müller-Spitzer, C., ur. (2014): *Using Online Dictionaries*. Berlin, Boston: De Gruyter Mouton.
- KOSEM, Iztok, et al. The image of the monolingual dictionary across Europe, Results of the European survey of dictionary use and culture. *International journal of lexicography*, ISSN 0950-3846, 2019, vol. 32, no. 1, str. 92-114, doi: 10.1093/ijl/ecy022.
- ARHAR HOLDT, Špela, KOSEM, Iztok, GANTAR, Polona, XVII EURALEX International Congress, 6-10 September, 2016, Tbilisi. Dictionary user typology : the Slovenian case. V: MARGALITADZE, Tinatin (ur.), MELADZE, George (ur.). Lexicography and linguistic diversity : proceedings of the XVII EURALEX International Congress, 6-10 September, 2016. Tbilisi: Ivane Javakhishvili Tbilisi State University. cop. 2016, str. 179-187