Identifying dictionary needs of language users by analysing user-generated content in digital media

Špela Arhar Holdt[1,2], Jaka Čibej[1], Ana Zwitter Vitez[1,3], Polona Gantar[1], Vojko Gorjanc[1]

[1] Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia
[2] Trojina, Institute for Applied Slovene Studies, Ljubljana, Slovenia
[3] Faculty of Humanities, University of Primorska, Koper, Slovenia
E-mail: spela.arhar@trojina.si, jaka.cibej@ff.uni-lj.si, ana.zwitter@guest.arnes.si, apolonija.gantar@guest.arnes.si, vojko.gorjanc@ff.uni-lj.si

1. Introduction

In Europe, research into dictionary use and user needs, expectations and abilities has been conducted for over 50 years. Common methodological approaches include questionnaires, interviews, tests, experiments, and different approaches to researching actual dictionary use such as dictionary protocols, eye-tracking, and log-file analysis (Welker 2013a, 2013b). The existing methods provide answers to several important questions and this information is an invaluable foundation for dictionary development and is being increasingly frequently used in lexicographical projects.

Until now, dictionary user research has focused on the user primarily from the point at which he or she actually uses the dictionary. The shortcoming of such a view is that it offers no perspective on the user's communication dilemma, which the dictionary should help resolve. As argued in literature (e.g. Bergenholtz and Tarp 2004; Tarp 2009) shifting focus to the latter might provide a better understanding of language users' needs, and consequently facilitate innovation in the presentation of lexical data to the user.

Some approaches to investigate extra- or pre-lexicographical situations have already been suggested, however, they are mostly qualitative as well as time-consuming and expensive (Tarp 2009: 293). In this paper, we present an alternative approach to identifying user needs, namely by examination language-related questions and comments posted on language advice sites, social media groups, and news forums.

2. Description of the study

Firstly, it is important to note that the approach we propose in this paper explores only those instances where the need for language information was recognized and publicly posted by the language users. These users are not perceived as 'dictionary users', as

it cannot be established if or how they actually use dictionaries or other language resources. Secondly, their background is in many instances unknown, which begs the question whether collecting a representative sample is at all possible. Thirdly, another unknown aspect is the extent of their activities as users of a certain type of media in comparison to other users or other media. Due to all these reasons, we cannot expect that this type of research could provide a representative picture of the entire populations' needs and opinions. However, specific samples can reveal a part of the users' needs, which can be taken as a starting point for a comparison with findings obtained with other methods.

The material for our study consists of 1209 language-related questions and comments collected from different digital sources, categorized and analysed as part of the preparation of a new monolingual dictionary of Slovene (Gorjanc et al., 2015). The results focusing mainly on the findings directly applicable to the monolingual dictionary project in question were published in Arhar Holdt et al. (2015) and Čibej et al. (2015), while a more methodologically oriented discussion is in preparation (Arhar Holdt et al.).

The questions and comments were collected from four different sources in terms of professionalism of the addressee, format, and communication medium: (I) a language advice website *ŠUSS* where language experts provide extensive answers to user-generated language questions; (II) a call-in radio show *Language Advice Service,* aired monthly (2009–2013) on Radio Slovenija 1, aimed at resolving listeners' language-related dilemmas; (III) three different Facebook groups, dealing with language-related dilemmas concerning translation, orthography and stylistic problems, and (IV) news forum sites with articles covering a debate about the ongoing plans for a new monolingual dictionary of Slovene.

The data for the analysis was prepared in two steps. First, the questions and comments were collected and arranged in spreadsheets. Then, we analysed the collected material in terms of its content, identifying and annotating the main problem category of each question or comment, as shown in Table 1.

	Question/Comment (translated from Slovene)	**Source**	**Category**
1	Lately I've been noticing the usage of the word *mnenjedajalec*, which I think is repulsive. Can you tell me what kind of word this actually is?? Are we allowed to use it??	ŠUSS	Is this word correct or not?
2	What is the difference between the words ***estetičen*** and *estetski*? I looked them both up in the Dictionary of Slovene Language, but they seem to be interchangeable.	Facebook	What is the difference between these words?
3	I'm preparing some technical documentation in a kindergarten and would like to know how to spell this word: *trakoder, trokoder, trokodero* or *trakader.*	ŠUSS	Which of these options is better?
4	[…] the above-mentioned dinosaur institutions should	News	Lexicographical

	finance their hazy projects with their own profits (and cover their own losses, too)	forums	institutions are exploiting taxpayers.

Table 1: Examples of language-related questions and comments

The quantitative analysis also aimed to identify additional relevant information: the user's attitude towards a certain language dilemma, the resources they attempted to find a solution in, the motivation behind the query, the formulation of the specific problem, and so forth.

3. Results and discussion

Due to the limited scope of this paper, we focus on the main findings regarding the user-generated language questions (I-III), leaving aside the comments on the news forums (IV), which are not discussing language problems as such but showing users' opinions on specific language policy decisions. A different approach is also required for questions at some of the Facebook groups, especially those targeted at professional translators who discuss language problems they encounter at their work. The quantitative results of the categorisation are presented in Figure 1.

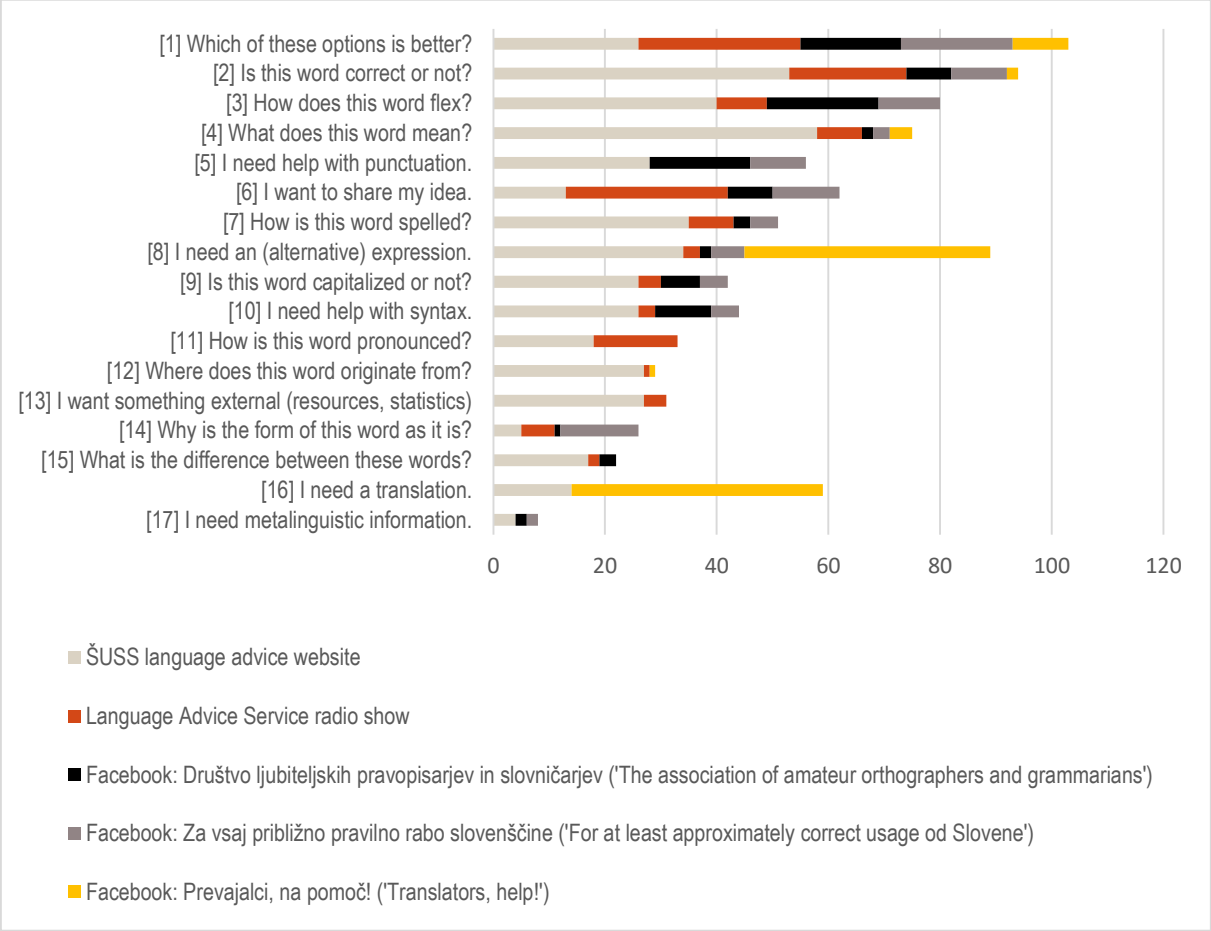

Figure 1: Language-related questions according to the expressed dilemma

An overview of the most common categories of user dilemmas provides guidelines in terms of content prioritization, as well as the structure and functionality of the dictionary interface. Solutions for many of the needs revealed by the analysed material have already been proposed, for example query lemmatisation, the did-you-mean function, pronunciation sound clips and interconnectivity with other resources, to list only a few mentioned in Lew and de Schryver (2014).

In addition, the data shows that users often wish to compare two or more language variants or expressive options. A digital dictionary could enable such a comparison by introducing parallel search and a display of more than one entry at a time. Secondly, the users demonstrate the need to express their own opinion, as previously shown by Müller Spitzer (2014: 156–159). This need can be addressed by creating a dictionary interface that facilitates discussion among the users as well as between the users and dictionary creators.

Furthermore, the analysis has revealed a number of user needs that go beyond a mere search for missing language information. For example, the users demonstrate the need for a deeper understanding of the solution to their dilemmas, as in many cases they expect not only an answer, but also an explanation of the rationale behind that answer. The digital medium enables us to address these needs by linking specific parts of dictionary content to the relevant language rules or explanations. Our study also identified users' need to evaluate or even judge the language use of other members of the language community. In terms of lexicography, this need is somewhat difficult to meet, and warrants not so much a reflection on how to further improve the dictionary, but rather how to prevent its misuse or abuse.

The analysed questions provide an insight into the background of the users' queries, their attitude towards particular language dilemmas, and the resources they use to find solutions. In some cases, the users explain the background of their query up to the length of a paragraph. Not all questions are this exhaustive, but the ones that are provide valuable authentic scenarios of language-related disruptions in which a dictionary could be consulted.

4. Conclusion

User-generated questions and comments reveal authentic problems as perceived and formulated by language users themselves as genuine responses to specific life situations. In this way, the approach proposed in this paper offers a more objective perspective on dictionary use compared to methods in which users report their problems post festum (e.g. interviews, questionnaires). On the qualitative level, the data shows – to a certain extent – the motivation behind the language question, and by that, the motivation behind a potential dictionary query. From this perspective, the method can complement e.g. log file analyses which reveal in great detail what people search for, yet in most cases cannot explain the reasons behind the searches.

With certain adaptations, the proposed method could be applied to other languages as well as to other activities aimed at providing language resources, tools, and services. It might be relevant to some readers that developing the concept of the research (along with data preparation) demanded approximately a month of research time (for three researchers). It also took another month to conduct the analyses and summarise the findings. From this perspective, the proposed method is relatively manageable in terms of time, even at this initial stage. However, with the development of (semi-)automatic procedures for data collection and categorization, a larger quantity of data could be processed in given time, which would further improve the reliability of the results.

5. Acknowledgements

The research was conducted at the Centre for Applied Linguistics (Trojina), Centre for Language Resources and Technologies (the University of Ljubljana), and was partially financed by the research programme “Slovene Language - Basic, Contrastive, and Applied Studies” (also at the University of Ljubljana). The research is partially based on the results of the Communication in Slovene project.

6. References

A. Data sources (Accessed on 23 December 2015)

Delo. http://www.delo.si/.
Dnevnik. https://www.dnevnik.si/.
Družina. http://www.druzina.si/.
Facebook group Društvo ljubiteljskih pravopisarjev in slovničarjev. https://www.facebook.com/groups/191388157545784/.
Facebook group Prevajalci, na pomoč! https://www.facebook.com/groups/help.prevajalci/.
Facebook group Za vsaj približno pravilno rabo slovenščine. https://www.facebook.com/groups/398216690214010/.
Jezikovni svetovalni servis. http://www.rtvslo.si/podcasts/svetovalni_serivs.xml.
RTV Slovenija. http://www.rtvslo.si/.
ŠUSS. http://www2.arnes.si/~lmarus/suss/index.html.

B. Other literature

Arhar Holdt, Š., Čibej, J. & Zwitter Vitez, A. (in press). Value of language-related questions and comments in digital media for lexicographical user research. *International Journal of Lexicography.*

Arhar Holdt, Š., Čibej, J. & Zwitter Vitez, A. (2015). S pomočjo uporabniških jezikovnih vprašanj in mnenj do boljšega slovarja. In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds.) *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL.

Bergenholtz, H. & Tarp, S. (2004). The Concept of Dictionary Usage. *Nordic Journal of English Studies* 3.1, pp. 23–36.

Čibej, J., Gorjanc, V. & Popič, D. (2015). Vloga jezikovnih vprašanj prevajalcev pri načrtovanju novega enojezičnega slovarja.' In V. Gorjanc, P. Gantar, I. Kosem & S. Krek (eds.) *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL.

Gorjanc, V., Gantar, P, Kosem, I. & Krek, S. (eds.) (2015). *Slovar sodobne slovenščine: problemi in rešitve.* Ljubljana: Znanstvena založba Filozofske fakultete UL.

Lew, R. & De Schryver, G.-M. (2014). Dictionary Users in the Digital Revolution. *International Journal of Lexicography* 27.4, pp. 341–359. doi:10.1093/ijl/ecu011.

Müller-Spitzer, C. (ed) (2014). *Using Online Dictionaries.* Berlin, Boston: De Gruyter Mouton.

Tarp, S. (2009). Reflections on Lexicographical User Research. *Lexikos* 19.1, pp. 275–296.

Welker, H. A. (2013a). Methods in Research of Dictionary Use. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* Berlin, New York: Walter de Gruyter, pp. 540–547.

Welker, H. A. (2013b). Empirical Research into Dictionary Use since 1990. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (eds.) *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* Berlin, New York: Walter de Gruyter, pp. 531–540.