Researching Dictionary Needs of Language Users Through Social Media: A Semi-Automatic Approach

Jaka Čibej[1,3]***, Špela Arhar Holdt***[2,3]

[1]Jožef Stefan Institute, [2]Faculty of Computer and Information Science, University of Ljubljana

[3]Centre for Language Resources and Technologies, University of Ljubljana

E-mail: jaka.cibej@ijs.si, spela.arharholdt@fri.uni-lj.si

Abstract

With the rise of digital media in the last decades, many language-related discussions have found home on various fora and social media such as Facebook, where users can participate in a shared-interest group to discuss language use, problems and resources. The posts in these groups are formulated by language users as a genuine response to a specific disruption in language use and offer an empirical starting point for studying language problems. We propose an automatic approach to extracting questions from language-related Facebook groups and describe the procedure in consecutive steps. We also address the issues of copyright, privacy and ethical constraints, and propose ways to overcome them. We present the extraction method on a case of two Slovene language-related Facebook groups: *Za vsaj približno pravilno rabo slovenščine* and *Društvo ljubiteljskih pravopisarjev in slovničarjev*. Both groups allow users to discuss language-related problems and find answers to their questions within the community. Our first extraction from these groups yielded approximately 1,900 posts (written by approximately 500 users) and 13,000 comments (posted by more than 900 users), providing ample material that can be analyzed to reveal the users' most frequent language problems.

Keywords: lexicographical user research, language problems, social media, automatic extraction, Facebook, Slovene

1 Introduction

As in many other fields, the development of the digital medium has brought an array of new possibilities to the field of dictionary user research. New procedures and methods used in this field, e.g. surveys, tests, evaluations, log file analyses (Welker 2013a, 2013b), have become less cumbersome or, in some cases, possible for the first time. This has enabled researchers to harness the change in interpersonal communication caused by the online environment. With the rise of digital media and computer-mediated communication in the last decades, many language-related discussions have found home on various fora and social media such as Facebook, where users can participate in a shared-interest group to discuss language use, problems and resources. The posts in these groups are formulated by language users as a genuine response to a specific disruption in language use. This data is especially valuable when taking into account the difference between what users believe their needs are (either in general, in relation to a specific language resource that is being evaluated, or when presented with hypothetical scenarios, which do not necessarily reflect their actual language dilemmas) and the actual language problems they encounter (i.e. what users really need when faced with an authentic language problem). From this perspective, observing user-reported language problems offers a more objective perspective on language problems compared to methods based on users reporting their problems post festum (e.g. interviews and questionnaires). Another aspect of this method that is of particular importance for user research is the broad scope of participants: while the population of Facebook cannot be considered as truly representative of all language users, the posts nevertheless reveal the problems, needs and opinions of a large and diverse number of language users, regardless of which language resources – if any – they use.

As we point out in the following sections, the method of collecting, classifying and conducting both a quantitative and qualitative analysis of self-reported language problems has already been tested. However, manual data extraction remains time-consuming and less than trivial. In this paper, we further develop this method by presenting a number of (semi-)automatic improvements. We first present an overview of related work done in this field and continue with a step-by-step description of the method to automatically extract data from Facebook groups in order to obtain a large quantity of Facebook posts representing authentic language problems, which can be analysed in order to obtain an overview of the most typical user needs. The main purpose of this approach is to facilitate the acquisition of empirical data on language users' authentic communication dilemmas, which the dictionary as a tool (alongside other language resources) should be designed to resolve. While we focus predominantly on Slovene data, the methodology is language-independent, as similar language-related discussion groups can be found for Slovene (*Za vsaj približno pravilno rabo slovenščine* 'For an at Least Approximately Correct Use of Slovene', *Društvo ljubiteljskih pravopisarjev in slovničarjev* 'Association of Amateur Orthographers and Grammarians'), Swedish (*Sverige mot särskrivning*, Sweden against Writing Separately; *Sprakpolisar*, 'Language Police'), Danish (*Sprog for sjov – og i alvor*, 'Language for fun – and for real'), Italian (*Gli amanti della lingua italiana* 'Fans of the Italian Language), and German (*Deutsch verbindet - Deutsch lernen* 'German Unites – Learning German'), to name just a few.

2 Related Work

In recent decades, lexicography has demonstrated an increasing interest in the needs, preferences and habits of dictionary users, with initiatives in dictionary-user research dating back as far as the 1960s (e.g. Barnhart 1962, Householder 1967, Tomaszczyk 1979) and gaining momentum in the 1980s (e.g. Hartman 1987, Wiegand 1987) and 1990s (e.g. Atkins 1998, Nesi 2000, Tono 2001). The emergence of the digital medium in the 2000s, however, allowed for new methodologies in dictionary-user research (Bergenholtz & Johnsen 2013, Müller-Spitzer 2014, Lew & De Schryver 2014). Different approaches – such as questionnaires, interviews, experiments, and research of actual dictionary use through think-aloud protocols, eye-tracking, log-file analysis, or user feedback collected through the dictionary interface – provide answers to which language resources dictionary users know and use, how they estimate their needs and habits in terms of the dictionaries they use, etc. This information is an invaluable foundation for dictionary development and has been increasingly frequently implemented in modern lexicographical projects.

However, existing approaches to dictionary users provide very little insight into why the user actually decided to consult the dictionary in question. Mentrup (1984: 160) proposed that the interest of the field '[. . .] should not start with the intangible dictionary usage situations but – as it were one level below – with language-related disruptions in language use situations'. This is later echoed by Tarp (2009), who suggests several possible approaches to address this gap, e.g. tests and interviews to investigate the readers' comprehension level and reception problems; analysis of text revisions; or simply the extension of existing methods (log files, eye-tracking, protocols) from dictionary use situations to extra-lexicographical situations, while already acknowledging that these approaches are mostly qualitative as well as time-consuming and expensive (ibid.: 293).

An alternative approach to identifying user needs, namely through user-generated content in digital media was proposed in Arhar Holdt et al. (2017) and Čibej et al. (2016). These two studies have confirmed that language-related discussions in social media groups provide a wide range of implicit and explicit information that can be useful when designing user-friendly and user-oriented language resources. However, both of them were based on a limited number of (at the time most recent) posts that were extracted manually, forming a small sample that was not representative of the entire group. The evaluation of the method highlighted that the procedure would benefit greatly from automatization.

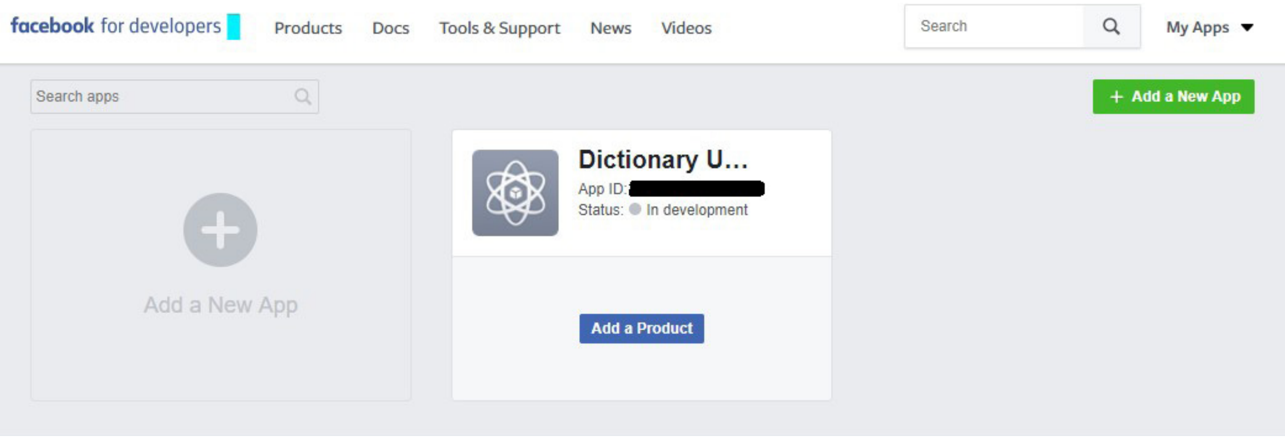
3 Automatic Extraction of User Posts

Automatic harvesting of information from social media is already commonplace in natural language processing (e.g. for sentiment analysis, opinion mining, and author profiling). We extend the use of this method to dictionary user studies by presenting an automatic approach to extracting questions from language-related Facebook groups through a Python script that makes use of the official Facebook Graph API. In this section and the following subsections, we describe the procedure in consecutive steps from identifying relevant Facebook groups and creating an app in the Facebook API, to extracting posts and comments. The Facebook Graph API allows for the extraction of all the posts (and comments posted as replies to those posts) from a Facebook group, along with a number of relevant metadata (e.g. user, time of publication, number of comments, likes and other reactions, links to resources and pages provided by the users) according to which a more representative and/or relevant sample can be made.

More detailed instructions on the use of the Python script are available on GitHub. In this paper, we only provide the basic steps.

3.1 Facebook Graph API

The Facebook Graph API is the primary way for apps to read (and write) to the Facebook social graph. In order to use the Facebook Graph API, a Facebook account is required. After logging in, the user must create an app with which to access Facebook data. The app must then be reviewed and approved by the Facebook staff, which usually takes several days.



Dictionary U...

App ID: [redacted]

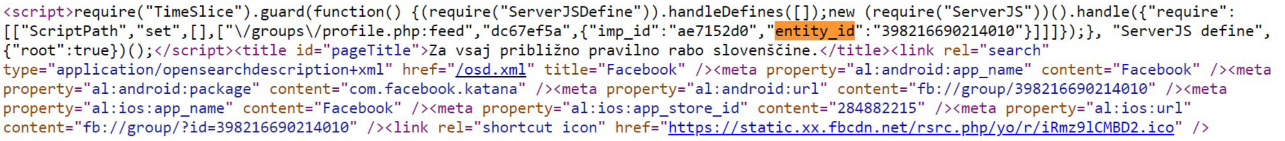
Status: ● In development

Figure 1: A screenshot of the Facebook Graph API interface. The censored part contains the app ID, which should be kept secret.

The user must then obtain the *app ID* (a unique, 16-digit number identifying the app), the *app token*, and the *access token* (an automatically generated sequence of alphanumeric characters that acts as an access code; a short-term access token expires in an hour, while an extended access token last for several months), all of which must be incorporated into the script.

3.2 Facebook group IDs and Access Permissions

Extracting data from a Facebook group also requires the group's ID number. As of April 2018, the ID of a group can be obtained by inspecting the source code of the group page and finding the 'entity_id' attribute (see Figure 2), which is a 15-digit number.



```
<script>require("TimeSlice").guard(function() {(require("ServerJSDefine")).handleDefines([]);new (require("ServerJS"))().handle({"require":
[["ScriptPath","set",[],["\/groups\/profile.php:feed","dc67ef5a",{"imp_id":"ae7152d0","entity_id":"398216690214010"}]]]});}, "ServerJS define",
{"root":true})();</script><title id="pageTitle">Za vsaj približno pravilno rabo slovenščine.</title><link rel="search"
type="application/opensearchdescription+xml" href="/osd.xml" title="Facebook" /><meta property="al:android:app_name" content="Facebook" /><meta
property="al:android:package" content="com.facebook.katana" /><meta property="al:android:url" content="fb://group/398216690214010" /><meta
property="al:ios:app_name" content="Facebook" /><meta property="al:ios:app_store_id" content="284882215" /><meta property="al:ios:url"
content="fb://group/?id=398216690214010" /><link rel="shortcut icon" href="https://static.xx.fbcdn.net/rsrc.php/yo/r/iRmz9lCMBD2.ico" />
```

Figure 2: A screenshot of the source code for the page of the public Slovene Facebook group *Za vsaj približno pravilno rabo slovenščine*. The group ID attribute is highlighted.

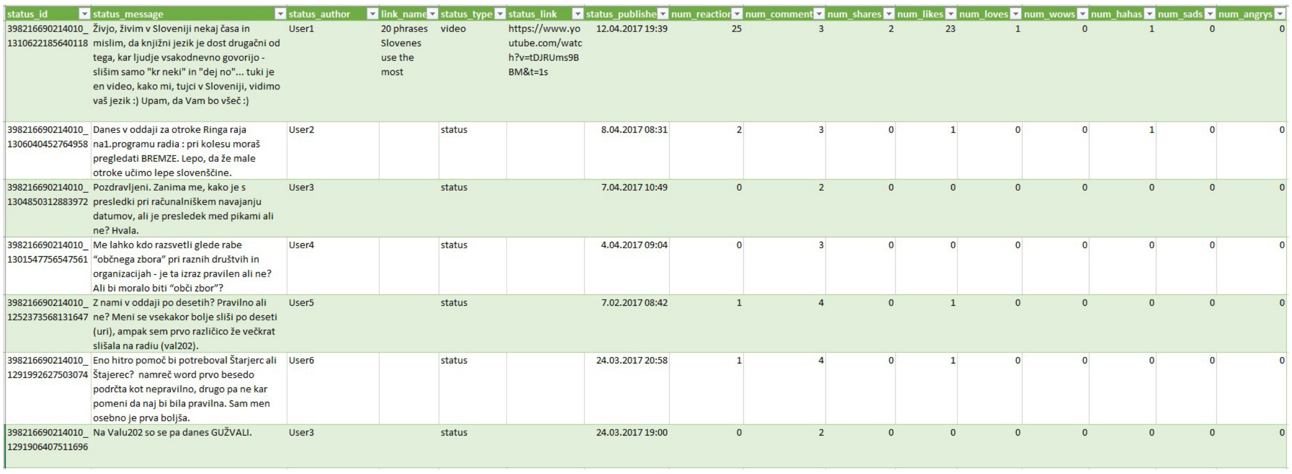
At this point, however, it should be noted that not all Facebook groups are equally accessible, as there are currently three types of groups: public groups (which new users can join freely once they have been confirmed by an administrator or another group member), closed groups (which are visible to the public, but can only be accessed by group members), and secret groups (which are invisible to the public and cannot be accessed by non-members). In terms of our post extraction method, there is an important difference between public and non-public groups. Data from public groups can be extracted by anyone with a Facebook account (and a Facebook Graph API app), regardless of whether they are a group member. With private (and secret) groups, however, this can only be done by the group administrator(s). The easiest way to bypass this restriction is to contact the group administrator(s) directly and explain the scope and goal of the research at hand. We discuss this in further detail in Section 5.

3.3 Data Extraction

The script we use[1] requires Python 3 and, in its current version (1.0), consists of two parts: the first part extracts the group posts, while the second extracts the comments beneath the posts. The output files can later be joined to form threads as seen in Facebook groups, or analysed separately – while user posts provide insight into the most frequent questions (or types of questions), user comments provide replies and, perhaps more importantly, the types of resources used to find solutions to the questions.

The script uses the Facebook Graph API syntax to request the following data: post ID, message (the text of the post or comment), username or pseudonym (e.g. User001), link (e.g. a link to a website if included in the post by the user), type of post (regular text-based post, link to video, etc.), time of publication, image (if included in the post by the user; the image is downloaded separately and is not included in the output file, but can be collocated with the correct post or comment through its ID), and finally, the number of comments, shares and different reactions (currently, Facebook allows users to mark posts with the following reactions: *like*, *love*, *wow*, *sad*, and *angry*). The data returned by the Graph API is first loaded in JSON format and then written to a CSV output file, which can be opened and analysed in most statistical analysis software (Excel, R, etc.). An example is shown in Figure 3.

[1] Our script is based on a script made by GitHub user Max Woolf (https://github.com/minimaxir/facebook-page-post-scraper). Our version is also available on GitHub: https://github.com/jakacibej/dictionary_user_needs



status_id	status_message	status_author	link_name	status_type	status_link	status_publishe	num_reactior	num_comment	num_shares	num_likes	num_loves	num_wows	num_hahas	num_sads	num_angrys
398216690214010_131062218564011 8	Živjo, živim v Sloveniji nekaj časa in mislim, da knjižni jezik je dost drugačni od tega, kar ljudje vsakodnevno govorijo - slišim samo "kr neki" in "dej no"... tuki je en video, kako mi, tujci v Sloveniji, vidimo vaš jezik :) Upam, da Vam bo všeč :)	User1	20 phrases Slovenes use the most	video	https://www.youtube.com/watch?v=tDJRUms9BBM&t=1s	12.04.2017 19:39	25	3	2	23	1	0	1	0	0
398216690214010_1306040452764958	Danes v oddaji za otroke Ringa raja na1.programu radia : pri kolesu moraš pregledati BREMZE. Lepo, da že male otroke učimo lepe slovenščine.	User2		status		8.04.2017 08:31	2	3	0	1	0	0	1	0	0
398216690214010_1304850312883972	Pozdravljeni. Zanima me, kako je s presledki pri računalniškem navajanju datumov, ali je presledek med pikami ali ne? Hvala.	User3		status		7.04.2017 10:49	0	2	0	0	0	0	0	0	0
398216690214010_1301547756547561	Me lahko kdo razsvetli glede rabe "občnega zbora" pri raznih društvih in organizacijah - je ta izraz pravilen ali ne? Ali bi moralo biti "obči zbor"?	User4		status		4.04.2017 09:04	0	3	0	0	0	0	0	0	0
398216690214010_1252373568131647	Z nami v oddaji po desetih? Pravilno ali ne? Meni se vsekakor bolje sliši po deseti (uri), ampak sem prvo različico že večkrat slišala na radiu (val202).	User5		status		7.02.2017 08:42	1	4	0	1	0	0	0	0	0
398216690214010_1291992627503074	Eno hitro pomoč bi potreboval Štarjerc ali Štajerec? namreč word prvo besedo podrčta kot nepravilno, drugo pa ne kar pomeni da naj bi bila pravilna. Sam men osebno je prva boljša.	User6		status		24.03.2017 20:58	1	4	0	1	0	0	0	0	0
398216690214010_1291906407511696	Na Valu202 so se pa danes GUŽVALI.	User3		status		24.03.2017 19:00	0	2	0	0	0	0	0	0	0

Figure 3: Example of an output CSV file imported into Excel.

By default, the script anonymizes all the usernames, replacing them with generic codes (*User1*, *User2*, etc.), which remove problems with privacy protection while still enabling posts to be grouped by user. The script does allow the automatic anonymization to be turned off, but in this case, researchers treat the extracted data as carefully as possible and take every precaution to protect user privacy (for instance, unanonymized data is not suitable for publication in publicly available corpora).

4 The Case of Slovene Language-Related Facebook Groups

We present the results of our automatic extraction method on a case of two Slovene language-related Facebook groups: *Za vsaj približno pravilno rabo slovenščine* (For an at Least Approximately Correct Use of Slovene) and *Društvo ljubiteljskih pravopisarjev in slovničarjev* (The Association of Amateur Orthographers and Grammarians). Both groups allow users to discuss language-related problems and find answers to their questions within the community.

4.1 Quantitative Overview

As of April 2018, the groups consist of more than 2,500 and 1,800 members, respectively, and have been active since 2011 and 2012, respectively. Our first extraction (see Table 1) from these groups yielded approximately 1,700 posts (written by approximately 500 users, some of which are members of both groups) and 13,000 comments (posted by more than 900 users). The data is shown in Table 1 below. As can be seen, the method provides ample material that can be analyzed to reveal the users' most frequent language problems and identify the areas in which existing language resources could be improved in order to better fulfil the needs of language users. We describe this in more detail in the following subsection (4.2).

Table 1: Number of users, posts and comments extracted from the groups.

Group	Users	Posts	Comments
Za vsaj približno pravilno rabo slovenščine	562	604	4.315
Društvo ljubiteljskih pravopisarjev in slovničarjev	273	1.135	8.548

An overview of the number of posts and comments per user shows that while the majority of users (approximately 90 %) posted only a handful of posts and comments (between 1 and 10), there are nevertheless several very productive users (with up to 105 posts and 822 comments). On average, users posted approximately 9 posts and 15 comments.

The fact that users post unevenly was one of the problems encountered with manual extraction of Facebook group posts. The sample collected in this way was very prone to skewing, as there is a higher chance to include only very active users while neglecting the ones that may have posted only a handful of questions, especially if they have not been very productive at the time of data collection. Automatic extraction allows for stratified sampling by user to ensure that all users that posted in the group are included in the sample.

4.2 Qualitative Overview

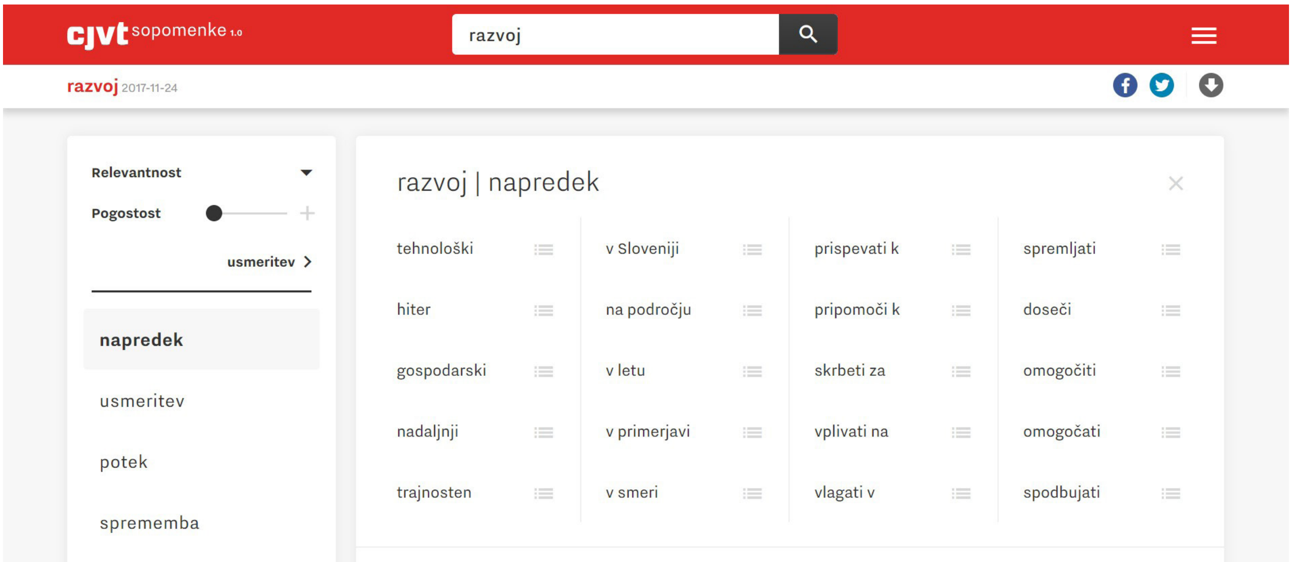
The posts are a valuable source of information to be implemented in the design of digital lexicographic resources, as already confirmed by the results of Arhar Holdt et al. (2017) and Čibej et al. (2016): according to their typology, the questions found in the posts can be divided into 17 categories, which cover diverse scenarios such as *Which of these options is better?*, *Is this word correct or not?*, *What does this word mean?*, and so on. The examples below are English translations[2] of posts extracted from the Facebook group *Za vsaj približno pravilno rabo slovenščine*. The questions cover a variety of different topics, including orthography (examples 4 and 5) and variation (examples 1 and 6), semantics, word form (example 2), word origin, translation (example 3), and metalinguistic or other external data.

(1) Hello. One question – šola astme (school of asthma), šola astma ali astma šola? And why. (I'm for 'šola astme' analogous to the expressions 'šola hujšanja' (school of weight loss), 'šola zdravega načina življenja' (school of healthy lifestyle), but I have no other arguments for it). Thanks. And have a nice Wednesday.
(2) How do we call the inhabitants of Sicily? (And I don't mean Italians ;))
(3) Does anyone know how to translate "zero anaphora"?
(4) UV light or UV-light?
(5) hi, I'd like to know how to correctly write the expression ad-hoc/ad hoc – in italics? (when speaking of an ad-hoc decision, an ad-hoc work group). thank you for the help&advice.
(6) when speaking of the Jedi from Star Wars: "jedijski" or "jedijevski" – the results in Gigafida are approximately equally frequent for both, with slightly greater frequency for the second. What do you think? Thanks for your replies.

The analysis and a thorough overview of the most common categories of user problems can provide a lexicographical project with several guidelines on how to prioritize dictionary content, how to structure the dictionary interface and what functionality it should offer. As pointed out by Arhar Holdt et al. (2017), for many of the needs revealed by the material extracted from language-related Facebook groups, a number of solutions are already available, for example query lemmatisation, the did-you-mean function, pronunciation sound clips, and interconnectivity with other resources (these are also mentioned in Lew and De Schryver (2014)). However, the analysis shows some user needs

[2] Slovene originals:
(1) Dan. Eno vprašanje - šola astme, šola astma ali astma šola? In zakaj. (Jaz zagovarjam 'šola astme' po vzoru šola hujšanja, šola zdravega načina življenja, drugega argumenta pa nimam). Hvala. In lep preostanek srede.
(2) Kako rečemo prebivalcem Sicilije? (In ne mislim Italijani ;))
(3) Morda kdo ve, kako se prevede "zero anaphora"?
(4) UV svetloba ali UV-svetloba?
(5) živijo. zanima me, kako pravilno zapišemo izraz ad-hoc/ad hoc - v italic? (ko govorimo o ad-hoc odločitvi, ad-hoc delovni skupini). hvala za pomoč&nasvet.
(6) v zvezi z jediji iz Vojn zvezd: "jedijski" ali "jedijevski"- Gigafida daje oboje v približno enakem številu, rahla prednost drugega. Kaj mislite? Hvala za odzive

that are not as frequently discussed, even though they could probably be met with relatively simple steps. For example, users often wish to compare two or more language variants. The comparison of two (semantically similar) words was also one of the most typical and frequent scenarios identified by Čibej et al. (2016), who analysed the posts in the Slovene Facebook group *Prevajalci, na pomoč!* (Translators, help!) and demonstrated that a great number of users would benefit from a Slovene synonym dictionary, a lacuna that has since been filled by the Thesaurus of Modern Slovene (Krek et al. 2017). The Thesaurus of Modern Slovene was designed as a direct response to the identified user needs, and among other functions, it offers the possibility of comparing two synonyms in context by providing their most typical collocates (see Figure 4, showing the most typical collocates for *razvoj* 'development' and *napredek* 'progress') and examples of use. This is thus a good-practice example of how the analysis of language-related user-generated content can directly contribute to user-friendly dictionary design.



razvoj | napredek

tehnološki	v Sloveniji	prispevati k	spremljati
hiter	na področju	pripomoči k	doseči
gospodarski	v letu	skrbeti za	omogočiti
nadaljnji	v primerjavi	vplivati na	omogočati
trajnosten	v smeri	vlagati v	spodbujati

Figure 4: Collocations page of the Thesaurus of Modern Slovene, allowing a comparison between two synonyms.

5 Personal Data Protection and Ethical Restrictions

When dealing with Facebook data, a number of legal and ethical restrictions need to be taken into account. In this section, we describe these issues and propose solutions to overcome them.

The first issue concerns personal data protection, as data obtained from Facebook most often contains personal information. In our case, the most problematic are the users' usernames, which usually consist of their real-life names and surnames. It is crucial to take every precaution to ensure that the users' rights to privacy are not violated. Our script automatically anonymizes all usernames, but also allows this option to be turned off (if names, and e.g. gender, which can be deduced from them, are important to the goals of the research at hand). In this case, the researcher(s) dealing with the data should ensure that all personal data is used only for research purposes and never shared outside the research group unless informed consent has been acquired from the group members and the material properly anonymized.

The second issue concerns ethical restrictions. In the case of public groups, the data and posts were publicly accessible and, until the most recent version of the Facebook Graph API (v2.12, April 2018), could be harvested even without explicit permissions from group members and/or administrators. Access has been restricted since then. In any case, it is advisable to establish contact with the group and explain the nature of the project, especially if the result (e.g. a language resource) will benefit the community. In the case of non-public groups, data has never been publicly accessible without the permission of the group administrator(s), so adequate contact with the community is obligatory. There are two ways a group administrator can grant access to the Facebook group data. The first way is by creating their own Facebook Graph API app and providing the researcher with a (temporary) access token that will allow the script to download group posts and comments. However, the access token either expires within an hour (which is usually too short a time to finish downloading all the data from the group) or within several months (which may raise suspicion among group members). In addition, this solution requires a lot of unnecessary work on the group administrator's part. The second way is to ask the group administrator(s) to accept the researcher's request to join the group and then temporarily (e.g. for a day or another fixed amount of time) promote them to group administrator. With administrator permissions, the researcher can then access the group's data through their own Facebook Graph API app. However, understandably, administrators will be reluctant to accept responsibility of allowing the data of the entire group to be accessed by a third party, which is why it is advisable to inform the community of the research taking place, the exact type and format of the data that will be collected, the purposes for which it will be used, and lastly, that data extraction will only take place at a pre-determined time, and anonymized. While it is often impossible or at least impractical to obtain consent from every single group user, a poll can be held within the group to vote on whether they are willing to allow access or not. The administrators can then determine a threshold, e.g. if more than 60 % of the votes are in favor of the data extraction, the researcher shall be granted access. It is also advisable for the researcher to draft an official statement signed by themselves and their institution, stating the conditions under which the data can be harvested (e.g. used only for scientific purposes).

Contacting a group is also important for dissemination purposes and community building. The researcher should keep in touch with the community even after data extraction to inform them about the progress of the project and perhaps post some interesting findings to allow the community to provide feedback. It is important not to treat the group simply as a source of information, but as a community that can contribute to dictionary design in a number of different stages of development.

6 Conclusion

In the paper, we have presented a method to automatically obtain large quantities of authentic language-related user questions (as well as their solutions) from Facebook groups on social media. The script used to extract posts and comments from Facebook groups is language-independent and is openly accessible on GitHub for the benefit of the research community. The extracted posts include invaluable implicit and explicit information that can be analysed in order to form guidelines for a more user-friendly and user-oriented approach to the design and compilation of new language resources. It is also worth noting that the method produces posts that include a number of relevant metadata that can be processed during the analysis to find or filter the most relevant posts, e.g. with the most comments or the most reactions. In addition, the method enables the creation of a sample that is more representative of the entire group, as it allows for stratified sampling by user.

However, the method does have several potential weak spots that need to be addressed. First, in light of recent controversial events with social media and discussions on data privacy, any restrictions to Facebook API policy, although unlikely to completely ban all automatic extraction, may prove problematic. Second, so far, it is impossible to extract metadata on the users themselves (e.g. their education level, age, etc.). While this method does sample a larger number of users compared to individual interviews, the results should not be interpreted as representative of the entire population of language users. In certain situations, groups include people with a shared professional background (e.g. *Prevajalci, na pomoč!* for translators), which allows the researcher to more accurately deduce the type of users being researched. In other situations, it might be prudent to conduct a poll within the group to determine, at least approximately, the type(s) of users being researched.

There are several other possibilities and improvements to the method to be explored as future work. First, within dictionary-compilation projects, it is possible to encourage the growth of a separate community to collect feedback on various versions of the project and, in later stages, to evaluate the interface. This method is being pioneered by the Thesaurus of Modern Slovene, which has a dedicated Facebook group aimed specifically at collecting user feedback on the Thesaurus. Feedback can then be automatically extracted and sorted by metadata.

Second, we intend to extract posts from all relevant Facebook groups for Slovene and conduct another analysis along the lines of Arhar Holdt et al. (2017), with particular emphasis on improving their bottom-up typology of user-generated language-related questions. Their analysis has namely shown that the method would benefit from a multi-layer categorization, with more robust categories for each layer if possible. These would be more adequate for further automatic processing. We namely also plan to implement machine learning to check whether user posts can be classified automatically, e.g. by language of interest (Slovene, English), by linguistic field (semantics, orthography, morphology, lexis), by potentially helpful resources (thesaurus, monolingual dictionary, bilingual dictionary), and so on. This will further automatize the entire process of analysing user needs through social media, and, if successful, provide an instant general overview of the most frequent user needs.

References

Arhar Holdt, Š., Čibej, J. & Zwitter Vitez, A. (2017). Value of language-related questions and comments in digital media for lexicographical user research. International journal of lexicography, 30 (3), pp. 285-308.

Atkins, B. T. S. (ed). 1998. *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*. Tübingen: Max Niemeyer Verlag.

Barnhart, C. L. (1962). Problems in Editing Commercial Monolingual Dictionaries. *International Journal of American Linguistics,* 28(2), pp. 161–181.

Bergenholtz, H. & Johnsen, M. (2013). User Research in the Field of Electronic Dictionaries: Methods, First Results, Proposals. 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. 556–568.

Čibej, J., Gorjanc, V. & Popič, D. (2016). XVII EURALEX International Congress, 6-10 September, 2016, Tbilisi. Analysing translators' language problems (and solutions) through user-generated content. In T. Margalitadze & G. Meladze (eds) Lexicography and linguistic diversity: proceedings of the XVII EURALEX International Congress. Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 158-167.

Householder, F. W. (1967). Summary Report. In F. W. Householder & S. Saporta (eds) *Problems in lexicography.* Bloomington: Indiana University Publications, pp. 279–282.

Krek, S., Laskowski, C. & Robnik-Šikonja, M. (2017). From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. In I. Kosem et al. (eds.) Proceedings of eLex 2017: Lexicography from Scratch, 19-21 September 2017. Leiden, Netherlands, pp. 93-109.

Lew, R. & De Schryver, G. M. (2014). Dictionary Users in the Digital Revolution. *International Journal of Lexicography,* 27(4), pp. 341–359.

Mentrup, W. (1984). 'Wörterbuchbenutzungssituationen–Sprachbenutzungssituationen. Anmerkungen Zur Verwendung Einiger Termini Bei HE Wiegand.' In W. Besch, K. Hufeland, V. Schupp & P. Wiehl (eds) *Festschrift für Siegfried Grosse zum 60. Geburtstag*. Göppingen: Kümmerle Verlag, pp. 143–173.

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

Nesi, H. (2000). *The Use and Abuse of EFL Dictionaries*. Tübingen: Max Niemeyer Verlag.

Tarp, S. (2009). Reflections on Lexicographical User Research. *Lexikos*, 19(1), pp. 275–296.

Tomaszczyk, J. (1979). Dictionaries: Users and Uses. *Glottodidactica* 12, pp. 103–119.

Tono, Y. (2001). *Research on Dictionary Use in the Context of Foreign Language Learning: Focus on Reading Comprehension.* Berlin: Walter de Gruyter.

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.

Acknowledgements

The research presented in this paper has received funding from the Ministry of Culture of the Republic of Slovenia (Thesaurus of Modern Slovene: By the Community for the Community), the infrastructural program of the Centre for Language Resources and Technologies (funded by ARRS, the Slovenian Research Agency) and the ARRS-funded research program P6-0215 Slovene language – basic, contrastive, and applied studies.