Univerza v Ljubljani

cjvt Centre for Language Resources and Technologies

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

Jaka Čibej[123], Špela Arhar Holdt[12]

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

Euralex, Ljubljana, 17. 7. 2018

# Background

- **Centre for Language Resources and Technologies, University of Ljubljana**
  - interdisciplinary research unit
  - systematic collection of empirical data on language user needs, habits and preferences
- Goals: openly-accessible and **user-friendly** language resources
  - address user needs as evidenced by empirical data
  - optimal solutions enabled by modern technology

# Researching user needs – Why social media?

- **Facebook** 
- valuable source of information
- commonly used in data mining and natural language processing
- also relevant for lexicography!
- language-related Facebook groups
  - communities of users discussing language and language problems
  - groups with different profiles (e.g. translators, language learners, grammar enthusiasts)

# Language-related Facebook groups

- Not a rare phenomenon!
- Groups available for a wide range of languages:
  - 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’ 
  - Sverige mot särskrivning, 'Sweden against Writing Separately' 
  - Sprakpolisar, 'Language Police' 
  - Sprog for sjov – og i alvor, 'Language for fun – and for real' 
  - Gli amanti della lingua italiana 'Fans of the Italian Language' 
  - Deutsch verbindet - Deutsch lernen 'German Unites – Learning German' 
  - ...

# User posts

question

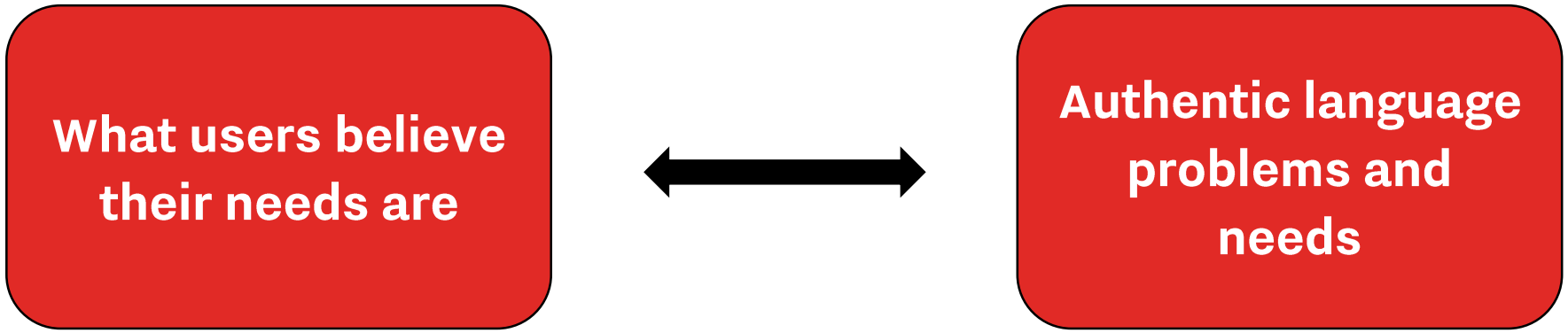
metadata
- publication time
- username

comments

reactions
- like
- wow/sad/...

# Why is this data relevant for researchers?

- digital approaches to dictionary user research have already been implemented:
  - online questionnaires, interviews, experiments
  - research of actual dictionary use through think-aloud protocols, eye-tracking, log-file analysis, or user feedback collected through the dictionary interface
- However – very little insight into **why** the user actually decided to consult the dictionary in question



# Advantages

- a more **objective** view
- **authentic** language problems (not hypothetical scenarios)
- **not limited** to a specific language resource
- **broad scope** of participants

# Does it work?

- First evaluations of the method showed **positive results** (Arhar Holdt et al. 2017, Čibej et al. 2016)
  - analysis of (manually collected) user posts in Slovene language-related Facebook groups
  - lots of implicit and explicit information – useful when designing user-friendly and user-oriented language resources

- **Room for improvement**
  - manual extraction is time-consuming
  - collecting additional metadata (available through the Facebook API)

- Next step – **automatic extraction**

# Automatic extraction of Facebook user posts

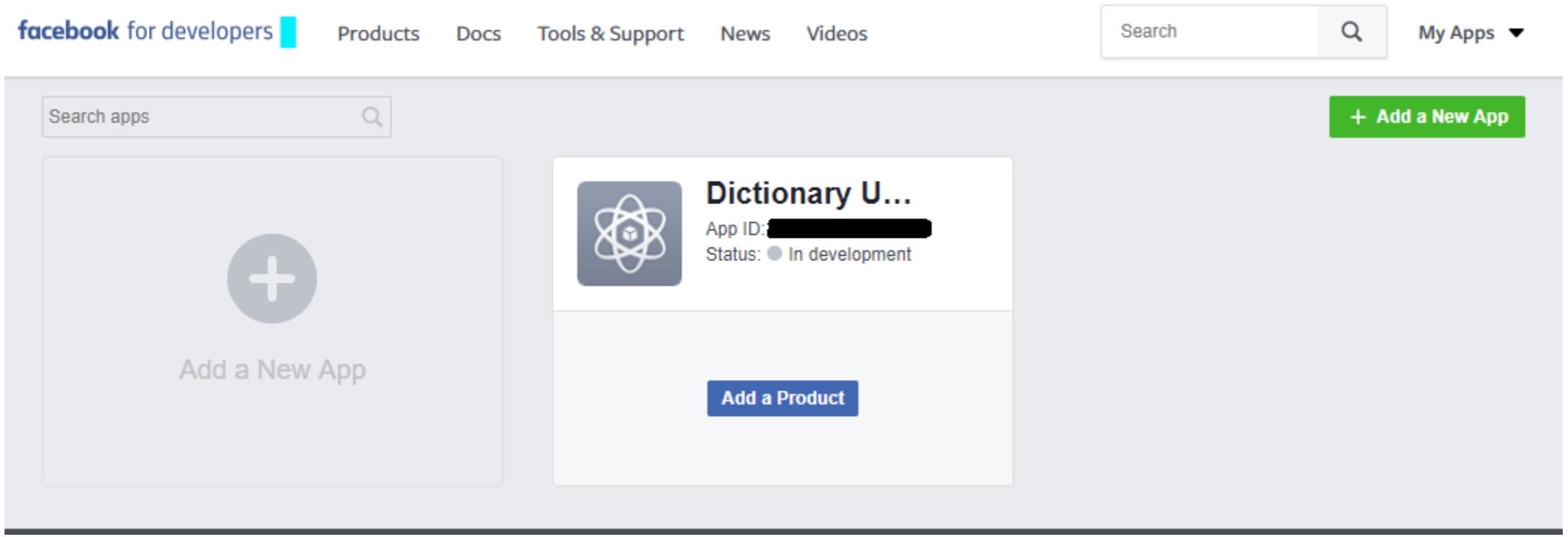
- **Python script**: https://github.com/jakacibej/dictionary_user_needs

- **Facebook Graph API**: https://developers.facebook.com/
- "the primary way for apps to read (and write) to the Facebook social graph"
- Free and (relatively) easy to use

- **Requirements**:
  - Facebook account
  - basic knowledge of the Facebook Graph API structure
  - Python 3 (to run our custom-made script)

# Step by step

- Step 1: Create a **Facebook app**, obtain the (secret) **app ID number** and (secret) **access token**.



# Step by step

- Step 2: Obtain the **ID number** of the relevant target group(s).
  - Inspecting the page source code and finding the 'entity_id‘ attribute.

<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" />

# Step by step

- Step 3: Obtain **administrator access** to the group.
  - Changes to the API since April 2018!
  - Accessing data requires administrator rights regardless of group status (public, closed or secret).

- **The simplest way**:
  - Contact the group administrator, explain the nature of your project and the data you intend to extract.
  - Inform the group and conduct a poll to determine whether members consent to being researched.
  - Obtain (temporary) administrator access and extract the data.
  - Keep the group informed and share the results of your research!
  - Respect data privacy!

# Extraction results

- CSV files containing post IDs, texts, (anonymized) usernames, links used, number of comments, shares, reactions, etc.

| 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_1310622185640118 | 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 |

# The case of Slovene groups

- *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
- more than 2,500 and 1,800 members; active since 2011/2012
- approx. 9 posts and 15 comments per user on average

| Group | Users | Posts | Comments |
|---|---|---|---|
| Za vsaj približno pravilno rabo slovenščine | 562 | 604 | 4,315 |
| Društvo ljubiteljskih pravopisarjev in slovničarjev | 273 | 1,135 | 8,548 |

# The case of Slovene groups

- user questions cover a variety of different topics: **orthography** and **variation**, **semantics**, **word form**, **word origin**, **translation**, and **metalinguistic or other external data**

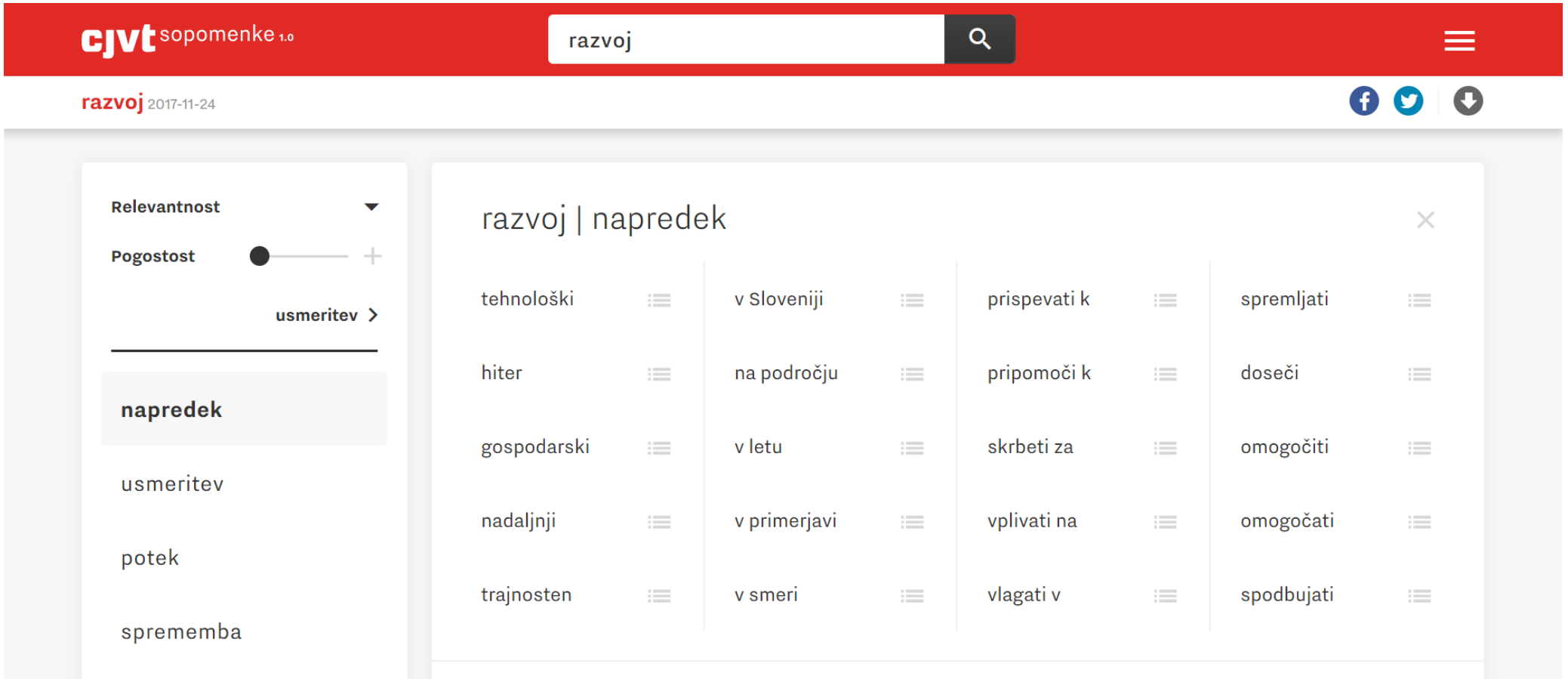
> Hello. One question – *šola astme* [school of asthma], *šola astma* ali *astma šola*? And why. […]

> UV light or UV-light?

> when speaking of the Jedi from Star Wars: "jedijski" or "jedijevski"? […]

# Case in point: The Thesaurus of Modern Slovene

- directly addresses the need to compare synonyms in context (Arhar Holdt et al. 2017, Čibej et al. 2016)



# Conclusion

1. Open-access script for automatic extraction of Facebook posts and comments from groups using the Facebook Graph API
2. Ample material for analysis
3. Data privacy and user involvement!
4. Future work #1: continued analysis of user posts
5. Future work #2: from automatic extraction to automatic classification?

Univerza v Ljubljani

cjvt Centre for Language Resources and Technologies

# Thank you!

Jaka Čibej[123], Špela Arhar Holdt[12]

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

# 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.