PROCEEDINGS OF SCIENCE

A different kind of dictionary – Collecting lexemes used in Austria together with citizens

Barbara Heinisch[a,*] **and Rebecca Stocker**[a] **and Esther Topitz**

aCentre for Translation Studies, University of Vienna,
Porzellangasse 4, Vienna, Austria

E-mail: barbara.heinisch@univie.ac.at, rebecca.stocker@univie.ac.at, esther.topitz@univie.ac.at

Lexicography is an area that is not only attractive for experts but also for amateur lexicographers who create dictionaries of their regional dialect(s). To build a bridge between the previously independent dictionary endeavours of experts and self-taught persons in lexicography, the citizen science dictionary *Wortgut* invites language communities to collect the variety of linguistic expressions used in Austria according to lexicographical principles. *Wortgut* does not focus on dialect alone but covers the entire spectrum of standard to non-standard German language. *Wortgut* allows citizens to collect linguistic expressions for every area of life. However, the researchers do not moderate the entries. Although the lexemes as entered by the participants can be subject of studies on language perception and attitude, the non-intervention makes it difficult to integrate them into other professional dictionaries. *Wortgut* helps people interested in language to preserve their local dialects, regiolects or sociolects and make them accessible to others.

Austrian Citizen Science Conference 2022 – ACSC 2022
28 - 30 June, 2022
Dornbirn, Austria

[*] Speaker

PoS(ACSC2022)006

1. Introduction: Citizen science and lexicography

Citizen science and language share the commonality that they are on everyone's lips. Therefore, the field of linguistics lends itself to be investigated by non-professional researchers. Since the field of linguistics is gaining ground in citizen science, this also relates to the motto of the Austrian Citizen Science Conference 2022, which investigates the factors of why citizen science should be conducted. This paper gives an example why lexicography can be an interesting field for citizen science.

Lexicography looks back on a long tradition of community-initiated or community-supported dictionary projects, e.g. volunteers create (dialect) dictionaries themselves or members of the public collect empirical evidence for local dialects on behalf of researchers. Moreover, there is an increasing number of projects in the field of linguistics on citizen science platforms or even dedicated platforms for citizen linguistics [1]. Additionally, citizens initiate their own dictionary projects through which they collect, for example, dialect lexemes in their region, which also serves to preserve their cultural heritage. All these initiatives demonstrate the interest of society in and the relevance of linguistic research as well as the (historically) strong connection between lexicography and citizen science.

1.1 Tradition of citizen linguistics for the German language

In German-speaking areas, the preparation of dialect dictionaries has traditionally been undertaken by the speakers themselves or in collaboration with speakers [2,3]. The Austrian saying that "every valley has its dialect" is also reflected by the popularity of dictionaries created by speakers of language varieties, especially speakers of dialects. Dictionaries may be created within associations dealing with the German language in Austria, including those that are addressing dialects as part of cultural heritage. These dictionaries more or less adhere to the standards in lexicography and are available in different formats, including searchable online dictionaries or glossaries in PDF format. This demonstrates that the coverage, findability and re-usability of these dictionary-like resources differs greatly.

Moreover, the exchange between so-called laypersons' dictionaries/folk dictionaries and professional dictionaries has been addressed only marginally. Therefore, the linguistic citizen science project *IamDiÖ – German in Austria* aims at bridging this gap and increasing the FAIRness (findability, accessibility, interoperability and re-use) of lexicographic resources produced by citizens. Finally, it also aims at upholding the tradition of lexicography conducted by citizens with the *Wortgut* online dictionary resource created by participants. Compared to traditional dictionaries, *Wortgut* is not restricted to a certain language variety, such as dialect, but covers the entire spectrum of standard and non-standard language.

1.2 The citizen science project 'German in Austria'

Wortgut is one of the initiatives of the citizen science project *IamDiÖ – German in Austria,* which invites members of the public to engage in linguistic research on the German language in Austria with activities like the Question of the Month [4], linguistic scavenger hunts [5] or meme contests.

2. The online *Wortgut* tool

Wortgut (lex.dioe.at) is an interactive citizen science dictionary tool, which is not created by lexicographers but by citizens. *Wortgut* is a newly developed tool aimed at bridging the gap between 'folk linguistics' and 'professional lexicography'. It focusses on standard and non-standard German language varieties in Austria. It was launched by IamDiÖ from the Centre for Translation Studies at the University of Vienna and funded by the Austrian Science Fund. *Wortgut* is not a comprehensive online dictionary, but it is constantly growing. The participants decide which expressions they would like to collect with *Wortgut* (Fig. 1) and how much data they add to their entry.

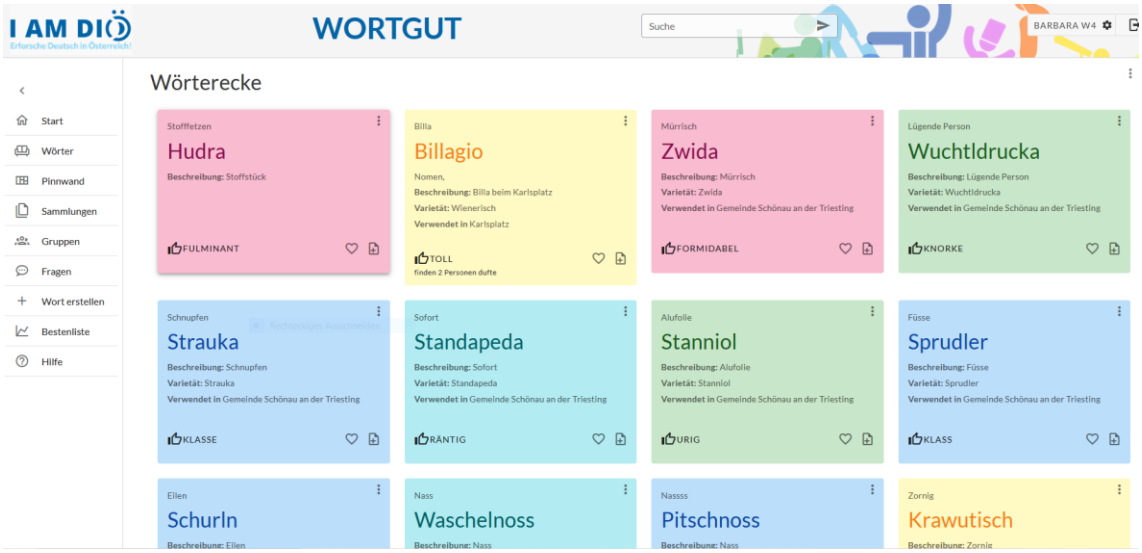

Figure 1: Online Wortgut dictionary tool: lex.dioe.at

By means of different web forms (Fig. 2), participants can enter their linguistic data according to lexicographical principles. There are only a few mandatory fields that participants have to complete when entering a lemma, such as a corresponding word from standard German or an explanation. However, they can provide various additional information about this lemma. Beginners usually only add the language variety, in which the lemma is being used and/or an example of use. More advanced users also add grammatical information, the area of life the lemma concerns or even the pronunciation according to the International Phonetic Alphabet or etymology. Moreover, participants can create so-called 'word collections' for a specific topic, e.g. sports or a regional dialect. They can work together on these word collections in groups.

The main contributors so far have been secondary school students as the website was launched as part of the Austrian Citizen Science Award in 2021. Students could choose their own topics and create word collections together. One class, for instance, focused on words that emerged in the COVID-19 pandemic, while another made a dictonary containing emojis, each showing the change in communciation due to changing circumstances.

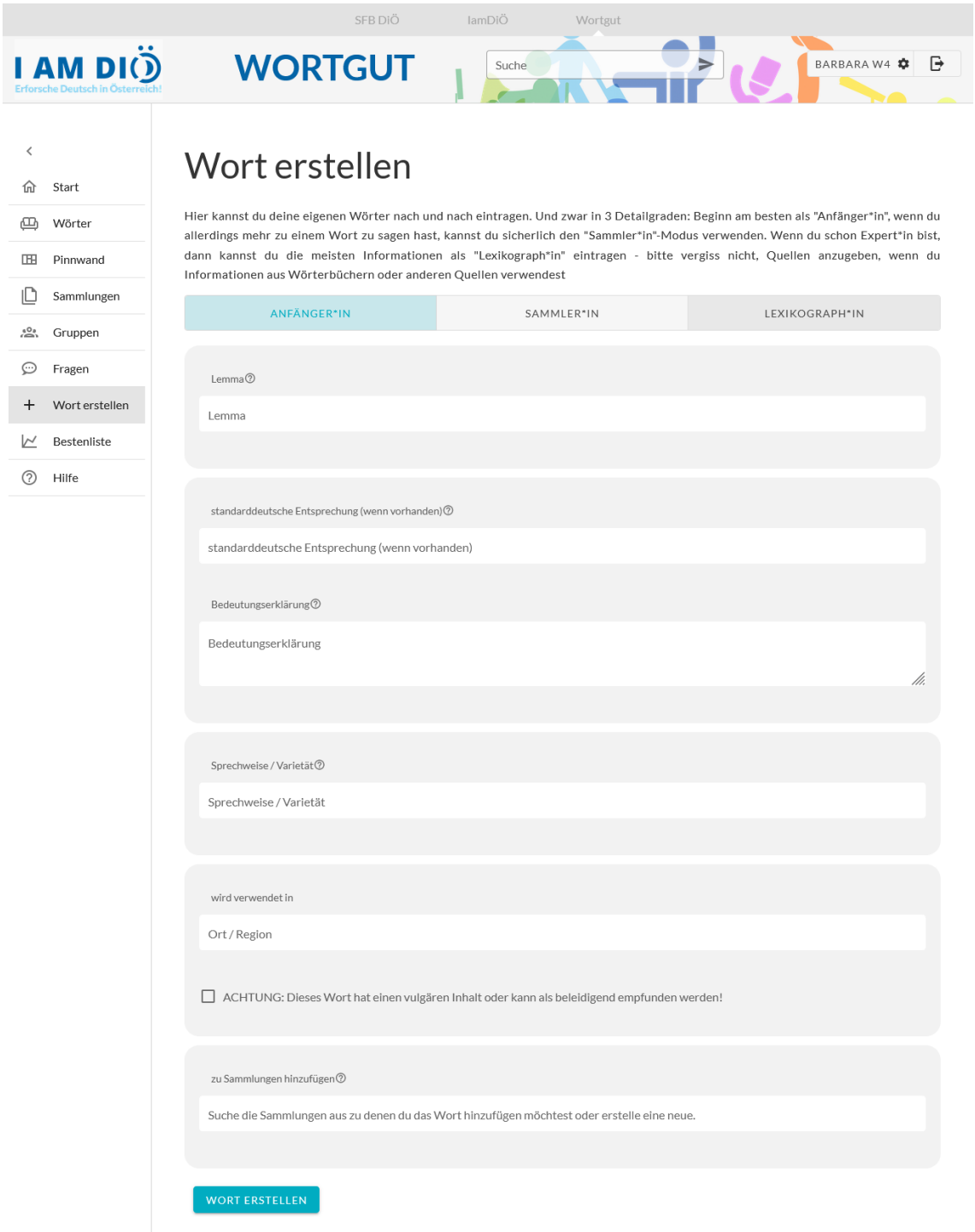

Figure 2: Forms for creating entries supporting different levels of expertise. In the screenshot: Form for beginners.

3.Engaging members of the public and data quality

In professional dictionary projects, lexicographers usually rely on the knowledge of the speakers of the language variety under investigation. While, in this case, professional lexicographers mainly use speakers as source for the compilation of a dictionary, folk dictionaries are often initiated and compiled by the speakers themselves. The latter were one of the target groups in the IamDiÖ project.

In a first step, the *Wortgut* dictionary tool was promoted among different target groups, including schools. All users can see the entries, but only registered ones can create entries in *Wortgut*. Users can only change their own entries, but they can comment and like the entries of other users or add them to their own word collections. Within the community, there is self-moderation. Users can tag their entries as 'offensive', for example, when adding swearwords. They can report discriminatory or offensive entries of other users to the researchers. Moreover, a comment function as well as a community forum allow users to discuss the meaning of entries or ask questions related to language. However, the community forum is only used to a limited extent.

Although workshops and guidelines on how to enter data in *Wortgut* were provided, not all entries meet the quality criteria. When a user enters their data, the tool does not check whether the entry meets the guidelines or whether the same lemma has already been entered by another user. This results in duplicate entries in *Wortgut*. Since it is easier to merge entries then to separate them retrospectively, duplicates can currently only be merged by the administrators.

Although this influences data quality, the researchers do not intervene in the participants' dictionary entries. Being too strict on data quality may repel users if they receive correction prompts from the system too often.

4. Outlook and conclusion

Due to the heterogeneity of the data in *Wortgut*, the next steps may include the introduction of a data validation loop to increase data quality, especially the adherence to lexicographical standards. However, the correctness of the data itself, e.g. whether a lemma is really used in a certain region and whether it has exactly the meaning as stated by the users, would require a larger and representative sample of speakers. Since *Wortgut*, as many other citizen science projects, relies on a self-selected group of participants, it is questionable whether this can be achieved through the current approach.

The participants were free to enter data of their choice. We did not restrict the subjects or topics. This openness in the project also seems to have stimulated the creativity of participants, who invented new ways in which the tool can be used. This can be exemplified by the emojis the participants entered as lemma and the explanation on the emojis' use. While this innovative and unintended use of the dictionary tool enriched the project, it also opens up questions about how to further use these data for research purposes.

To conclude, lexicography is an area that is not only interesting to experts but also to hobby lexicographers passionate about collecting regional dialects or youth language. To bridge the gap between official dictionaries and folk dictionaries, *Wortgut* invites the language communities to collect the entire range of linguistic diversity in Austria according to lexicographical principles, covering the entire spectrum of standard and non-standard language. Thus, *Wortgut* supports the preservation and visibility of linguistic expressions and gives an insight into specialised language, youth language, standard language, vernacular, dialect or even emojis.

The expected benefits for the speaker communities are self-expression (of their linguistic identity), raising awareness of linguistic diversity, including the diversity of dialects in Austria and the acknowledgement of dialect competence as a language competence. This should also help combat the stigmatisation of dialect speakers as being uneducated [6, 7] and reveal language change, for example in the case of youth language or through contact with languages of

immigrants. Therefore, *Wortgut* helps people interested in language to preserve their local dialects, regiolects or sociolects and make them accessible to others.

Acknowledgment: This research received funding from the Austrian Science Fund (FWF): TCS 57G. Conception of *Wortgut*: Ludwig M. Breuer, Rebecca Stocker, Esther Topitz, Barbara Heinisch. Implementation of *Wortgut*: Andreas Olschnögger. Feedback on *Wortgut*: Markus Pluschkovits.

References

[1] J. Fiumara, C. Cieri, J. Wright, and M. Liberman, *LanguageARC: Developing Language Resources Through Citizen Linguistics*, in: *CLLRD 2020*, 2020, pp. 1–6.

[2] DiWA, *Die Rolle des Wenker-Atlasses in der Geschichte der Dialektologie*. Available at http://www.diwa.info/Geschichte/RolleDesWenkeratlasses.aspx.

[3] P. Stöckle, *Wie ein Dialektwörterbuch entsteht: Wie kommen eigentlich die Wörter ins Wörterbuch? – Ein Bericht aus der WBÖ-Werkstatt*. Available at https://iam.dioe.at/blog/1963.

[4] B. Heinisch, *Reaching the limits of co-creation in citizen science — exemplified by the linguistic citizen humanities project 'On everyone's mind and lips — German in Austria',* JCOM 20 (2021), A05.

[5] B. Heinisch, *Hunting for signs in the public space – the method of linguistic treasure hunts as a form of citizen science*, in: *Conference Proceedings of the 5th Austrian Citizen Science Conference 2019 (ACSC2019)*: 26-28, June, 2019, Obergurgl, Austria. Proceedings of Science, 2020.

[6] K.J. Mattheier, *Dialekt und Standardsprache. Über das Varietätensystem des Deutschen in der Bundesrepublik*, in*: Int. J. Soc. Lang*, 1990.

[7] B. Soukup, *Über die empirische Spracheinstellungsforschung in Österreich,* in: *Akustische Phonetik und ihre multidisziplinären Aspekte. Ein Gedenkband für Sylvia Moosmüller,* 2022.