# Creating a Lexicon of Bavarian Dialect by Means of Facebook Language Data and Crowdsourcing

**Manuel Burghardt, Daniel Granvogl, Christian Wolff**

Media Informatics Group

Institute for Information and Media, Language and Culture

University of Regensburg, Germany

Email: manuel.burghardt@ur.de, daniel@granvogl.com, christian.wolff@ur.de

**Abstract**

Data acquisition in dialectology is typically a tedious task, as dialect samples of spoken language have to be collected via questionnaires or interviews. In this article, we suggest to use the "web as a corpus" approach for dialectology. We present a case study that demonstrates how authentic language data for the Bavarian dialect (ISO 639-3:bar) can be collected automatically from the social network Facebook. We also show that Facebook can be used effectively as a crowdsourcing platform, where users are willing to translate dialect words collaboratively in order to create a common lexicon of their Bavarian dialect. Key insights from the case study are summarized as "lessons learned", together with suggestions for future enhancements of the lexicon creation approach.

**Keywords:** dialectology, Bavarian, ISO 639-3:bar, dialect lexicon, crowdsourcing, social media, Facebook

## 1. Introduction: Dialectology and the Internet

Dialectology is a branch of sociolinguistics that typically examines instances of spoken language. This makes data acquisition a tedious task, as dialect samples of spoken language have to be collected via questionnaires or interviews. Furthermore, data collected in such a way needs to be transcribed and normalized. In corpus linguistics, there has been a trend to utilize the "web as a corpus" during the last years (Kilgarriff & Grefenstette, 2003; Baroni et al., 2009). More recently, there are also a number of studies from the field of dialectology that try to make use of language resources available from the Internet: For the case of Swiss dialect, Siebenhaar (2003) estimates that around 22% percent of Swiss websites contain text written in dialect. In a consecutive study on Swiss Internet Relay Chats (IRC), Siebenhaar (2005) finds that 80-90% of the messages posted in the chatrooms he analyzed are written dialect. Ziegler (2005) also presents a study on IRC chats, studying how German dialect is realized in chatrooms, and how it differs from standard language use. As IRC is quickly becoming outdated in the light of more recent social media platforms such as Twitter or Facebook, we believe that dialectologists should also try to make use of language data created by the users of these novel social networks.

In this article, we present a case study that demonstrates how authentic language data for a Bavarian dialect (ISO 639-3:bar) can be collected automatically from the social network Facebook. We also show that Facebook can be used effectively as a crowdsourcing platform, where users are willing to collaboratively translate dialect words in order to create a common lexicon of their Bavarian dialect.

## 2. Overview of Bavarian Dialect

Bavaria is one of 16 federal states in Germany. However, it is important to distinguish the state of Bavaria ("Freistaat Bayern") and the Bavarian dialect, which is not per se identical: Not all inhabitants of Bavaria speak Bavarian dialect, and there are also speakers of Bavarian dialect outside of Bavaria, e.g. in Austria or South Tyrol (Zehetner, 1985: 16). Furthermore, Bavaria is by no means a coherent dialect space (Zehetner, 2014:13). Accordingly, Zehetner (1985: 71) suggests a structuring of Bavarian dialects into 5 major dialect families that all can be distinguished by distinctive dialect features, and that can be associated with 10 different regions in Bavaria (cf. Table 1).

| **Dialect family** | **Regions in Bavaria** |
|---|---|
| Nordbairisch (*northern Bavarian*) | Nörd. Oberpfalz (*northern part of the Upper Palatinate*) / östl. Oberfranken (*eastern part of Upper Franconia*) |
| | Westl. Oberpfalz (*western part of the Upper Palatinate*)/ östl. Mittelfranken (*eastern part of Middle Franconia*) |
| | Mittlere Oberpfalz (*middle part of the Upper Palatinate*) |
| Nordmittelbairisch (*northern middle Bavarian*) | Südl. Oberpfalz (*southern part of the Upper Palatinate*)/ nördl. Niederbayern (*northern part of Lower Bavaria*) |
| | Mittlerer Bayerischer Wald (*middle part of the Bavarian forest*) |
| Mittelbairisch (*middle Bavarian*) | Unterer Bayerischer Wald (*lower part of the Bavarian forest*) |
| | Ober- und Niederbayern (*Upper and Lower Bavaria*) |
| | Westl. Oberbayern (*western part of Upper Bavaria*) |
| Südmittelbairisch (*southern middle Bavarian*) | Oberbayerisches Alpengebiet (*the alpine region of Upper Bavaria*) |
| Südbairisch (*southern Bavarian*) | Werdenfelser Land, Isarwinkel |

Table 1: Overview of the main Bavarian dialect families and the regions where they occur.

## 3. Corpus Creation from Facebook Language Data

One important goal of this case study was to create a corpus of dialect language data from a freely accessible, social media platform, such as Twitter or Facebook. We decided to use Facebook as a source for dialect data for the following reasons: In Twitter, users have to keep their messages short (maximum length of a Tweet: 140 characters); Facebook has no limitations with regard to the length of a message. We believe that the freedom to write messages without having to worry about length restrictions will result in a more 'natural' language usage that will be better suited for the collection of dialect samples. The main reason for choosing Facebook, however, is the availability of a large number of open, thematic groups that can be easily accessed via the Facebook Graph API[1]. While users in Facebook typically maintain a private profile in order to share content and communicate with people from their personal social network, groups are used to communicate with people who are not explicitly part of one's personal social network. Groups on Facebook are usually focused on a specific topic. There are open groups that can be joined by anybody, but also closed groups that are restricted to those users who are invited by the group's moderators. The availability of several open groups that are more or less explicitly dedicated to Bavarian dialect was the main reason and also the initial inspiration for this project. Many of these groups are dedicated to a specific city[2], but there are also groups that span larger regions (cf. Table 1). These groups typically consist of several hundred members who discuss and write about different topics in Bavarian dialect. For our case study, we decided to use the group "Niederborisch für Anfänger und Runaways" [3], which showed to have a very lively and active community with approx. 850 members, who regularly engage in discussions about regional peculiarities and subsequently write in dialect form. Another reason for choosing this group is that its "Mittelbairisch" dialect is the Bavarian dialect with the most speakers (Zehetner, 1985: 12).

In order to build a corpus from the messages posted in this group we have created a crawler that can be used to extract the message text of a group via the Facebook Graph API. For our case study, we created a corpus on May 9, 2014, which contains all messages posted since the creation of the group in 2009. The raw corpus contains 86,339 words (counted using VoyantTools [4]). After the elimination of emoticons and various special characters, we have created a database that contains one instance of every running word, the total number of occurrences of that word in the corpus, and the left and right context (10 words on each side) for the first occurrence of the word. The database comprises a total of 16,560 unique words. As the words in the database also contain numerous samples of Internet language, we filtered most of the non-dialect words by means of a custom stop words list that is based on a precompiled corpus of Internet language, the *Dortmund Chat-Korpus* (DCK) (Beißwenger, 2013). For our stop words list, we used a publicly available subcorpus of the DCK, which is called the "release corpus"[5]. In order to get rid of unwanted DCK metadata, such as timestamps or nicknames, we only used data from the actual message text that could be easily identified in the corpus via the XML tag "messageBody". As our goal was to create a stop word list from this corpus, we reduced the 212,835 tokens to a wordlist that eventually contains 24,422 unique word forms. The wordlist was created by means of the freely available AntConc[6] tool. Using this stop words list on our Facebook wordlist reduced the original 16,560 words to 13,466 words (cf. Fig. 1 for an overview of all those steps). For a more detailed discussion of this filtering step please cf. Section 8.1.

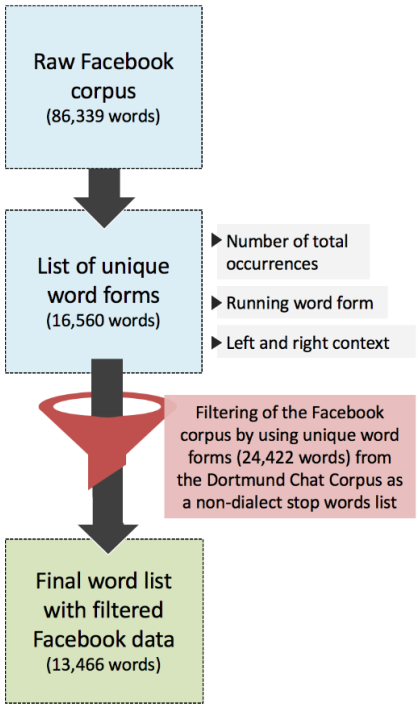


Figure 1: Basic steps in the creation of the final corpus, which is a filtered wordlist of the initial Facebook data.

## 4. Characterizing the Data

Examining the frequency distribution of the final word list, we found a typical Zipf distribution, i.e. very few words (only 28 of 13,466) occur with a frequency higher than 100, with the highest frequency being 495. The bulk of words occurs only once (9,814) or twice (1,711). Taking a closer

[1] https://developers.facebook.com/docs/graph-api; Note: all URLs referenced in this paper were last accessed on March 9, 2016.

[2] https://www.facebook.com/groups/ein.echter.chamer/; https://www.facebook.com/groups/echtestraubinger/

[3] https://www.facebook.com/groups/121572707986445/

[4] http://voyant-tools.org/

[5] http://www.chatkorpus.tu-dortmund.de/korpora.html#releasekorpus

[6] http://www.laurenceanthony.net/software.html

look at the word list, we observed that most of the words with a higher frequency (403 words occur >= 10 times) are actual dialect words. Among the few non-dialect words that occur are names of persons (11) or places (2), and one written emoticon (**lach** = laughing).

A snippet of the 20 most frequent dialect words (cf. Table 2) reveals similarities to frequent Standard German words, which include common pronouns, conjunctions, prepositions and particles, but also different variations of the auxiliary verb "haben" (*have*). The other words are examples for common adjectives, e.g. "gut" (*good*) and "schön" (*beautiful*), and adverbs, e.g. "wieder" (*again*), "heute" (*today*), "gerade" (*just*) and "einmal" (*once*). The data shows that the same dialect word can be expressed in a number of orthographic variations, e.g. "mei / mej" (*my*) or "oba / owa" (*but*), which is not surprising, as there is no standard orthography for Bavarian. At the same time, some of the words are obviously homographs, e.g. "grod". These issues, and the problems that emerge for the crowdsourced translation of these words, are discussed in some more detail in Section 8.2.

| *Rank* | *Word form* | *Standard German* | *POS* | *Freq.* |
|---|---|---|---|---|
| 01 | hod | hat (haben) | verb | 495 |
| 02 | hob | habe (haben) | verb | 389 |
| 03 | mei | mein | pronoun | 383 |
| 04 | oba | aber | particle | 317 |
| 05 | af | auf | preposition | 270 |
| 06 | wieda | wieder | adverb | 221 |
| 07 | mid | mit | preposition | 203 |
| 08 | guad | gut | adjective | 203 |
| 09 | hoid | halt | particle | 196 |
| 10 | heid | heute | adverb | 171 |
| 11 | oda | oder | conjunction | 158 |
| 12 | host | hast (haben) | verb | 155 |
| 13 | grod | gerade / Grad | adverb / noun | 151 |
| 14 | owa | aber | particle | 139 |
| 15 | mej | mein | pronoun | 136 |
| 16 | woas | weiß (wissen) | verb | 130 |
| 17 | ois | alles | pronoun | 125 |
| 18 | hosd | hast (haben) | verb | 125 |
| 19 | amoi | einmal | adverb | 124 |
| 20 | schee | schön | adjective | 122 |

Table 2: Overview of the 20 most frequent dialect word forms.

## 5. Example Word Formation Analyses

The corpus can be analyzed with regard to typical features of Bavarian dialect, as suggested by Zehetner (1985: 54ff; 143ff.). On the level of word formation, which can be examined very well with our written dialect corpus, a distinctive feature of Bavarian is the use of diminutive for nouns (sentences 1+2) and verbs (sentences 3+4).

(1) *... hod´s um 3 in da friah bei uns des **liedl** lautstark auf der strass gsunga.*
(2) *... da voda hot se meistens mehra drüber afgregt wia's **deandl**.*
(3) *... na, **zindln** dama ned, sama scho brav.*
(4) *... a bisse **rumstandln**, weil dsunn scheint so scheeee.*

Another feature of Bavarian dialect is concerned with the ending of adjectives, which is quite different from the standard German variant. Typical endings for Bavarian adjectives are "-ad" (sentences 5+6) and "-ig" (sentences 7+8).

(5) *...da lebatran hod wir a **stingad(a)** fisch gschmekt ...*
(6) *... haha, ok, is aweng **siaslad**.*
(7) *... daand de preissen den grünkohl ganz fett kocha und irgenda **greislig(e)** wurscht dazuaessn.*
(8) *... mir ist ein pfundiger preuße lieber als ein **grantig(er)** bayer.*

These example analyses illustrate that a Bavarian dialect corpus, gathered automatically from social media data, can be used to examine established categories of Bavarian dialectology.

## 6. Crowdsourced Lexicon Translation

Besides the creation of a dialect corpus from a freely accessible Facebook group, we also wanted to examine whether the community of users from which the dialect language data was collected, is willing to translate their own dialect words. For our crowdsourcing experiment we decided to use the 60 most frequent dialect words, and to have them translated by the members of the corresponding Facebook group. We designed a web tool that allows users to translate those selected words[7]. To keep the threshold for participation low, we did not implement an authentication mechanism, i.e. users were able to visit the translation site and start translating right away. In the tool interface, all words are presented to the user with their left and right context. Translations can be entered into an empty text field. If users are unable to translate a word, they may choose from the following two options: "I don't think this is a Bavarian expression at all" or "I don't know an adequate translation for this word". By clicking on the *save* button, the information is stored in a MySQL database and the next word is presented to the user. The order of words is randomized for every user. Users may stop translating at any time, i.e. they can translate as much as they want.

[7] The crowdsourcing tool can be experienced via http://bayerisch-deutsch.granivogl.de/home-uebersetzen/.

## 7. Translation Results

The link to the translation tool was posted in the Facebook group on August 12, 2014. Most of the visits occurred during the first 4 days; the whole experiment lasted for 10 days. In the end, 161 group members (total group size 848) visited the translation page and created a total of 3,655 translations. In most cases, there are multiple translation variants, but typically one variant is way more frequent than the other variants. That is why we compiled a lexicon with all possible translations (cf. Table 3), but also a version of the lexicon that only contains the most frequent translation for each dialect word.

| Term | Translation | Frequency |
|---|---|---|
| *af* | auf | 66 |
| | auf einmal | 1 |
| | not a Bavarian word | 1 |
| | no translation found | 1 |
| *amoi* | einmal | 49 |
| | auch mal | 5 |
| | ein mal | 3 |
| | auch einmal | 3 |
| | mal | 2 |
| | (ein)mal | 1 |
| | no translation found | 1 |
| | wiedermal | 1 |

Table 3: Two examples from the lexicon, with all potential translations ordered by frequency.

## 8. Lessons Learned and Future Directions

This section summarizes some key insights from our case study, and makes suggestions for future enhancements of the approach.

### 8.1 Data collection and filtering

Taking a closer look at the filtered corpus, it shows that a number of dialect words are homographs for German non-dialect words. The dialect meaning of "affe" (cf. sentence 9) is "hinauf" (*on*), whereas the non-dialect meaning is "Affe" (*monkey*). The dialect meaning of "nixe" (cf. sentence 10) is "nichts" (*nothing*), whereas the non-dialect meaning is "Nixe" (*mermaid*).

*(9) ... wer es a weng scherfa mog der konn pfeffa a no **affe** doa ...*
*(10) ... wieso sogsdan nochad do **nixe** ...*

Although the two examples were not excluded from the corpus, as they did not occur in the stop words list, there still is the danger that potential dialect words are lost because of such a filtering. As the filtering reduced the initial corpus by only 19%, we might be inclined to completely skip the filtering step for future studies, and rather rely on the crowd for tagging the most frequent words as being dialect or non-dialect.

### 8.2 Crowdsourced translation

A number of issues could be discovered during the experiment, in which the community translated their own dialect by means of a crowdsourcing tool:

- Some users entered multiple, comma-separated translation variants that had to be separated manually afterwards. The tool will be adapted in a way it allows users to enter multiple variants in different text fields.
- A lot of variation in the translations also came from orthographic ambiguities (examples: *muss* vs. *muß*, *zu hause* vs. *zuhause*, *täte* vs. *taete*, etc.). The tool will be adapted in a way to recognize such obvious ambiguities.
- The same is true for the dialect words that are to be translated (example: *eitz* vs. *ejz*); here, an automatic recognition of orthographic variation will be more difficult, as there is more spelling variation. As there is no standard orthography for Bavarian, we believe that it is important to collect different written manifestations of spoken dialect.
- When translating a verb, many users add a personal pronoun to their translation (example: *host* → *hast* vs. *hast du*). The tool will be adapted in a way to recognize personal pronouns in this type of scenario.
- Many users added an explanation into the translation text field. The tool will be adapted in a way to provide a separated commentary field, which facilitates the distinction of translation and comments or explanations.
- Some homographs (example: *grod* → *Grad* (*degree centigrade*) vs. *gerade* (*straight*)) were not translated properly, which indicates that users do not consider the context of the word appropriately. The tool will be adapted in a way to visualize the left and right context more prominently.

For the lexicon in our case study, we manually revised it according to the described lessons learned. We were able to reduce the initial 327 different translations for 60 dialect words to 233 translations.

## 9. Conclusion

In this article we have shown that Bavarian dialect data can be collected from dedicated Facebook groups. We believe the approach is also feasible for other German dialects, as there exists a great number of groups that can be matched with a specific dialect region. The participation rate (19%) of the group members for the collaborative translation was rather high, and produced a decent number of viable translations that allow for interesting insights into the use of dialect on the Internet.

## 10. References

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The wacky wide web: a collection of very large

linguistically processed web-crawled corpora. Language Resources and Evaluation, 43(3), pp. 209-226.
Beißwenger, M. (2013). Das Dortmunder Chat-Korpus. In: Zeitschrift für germanistische Linguistik 41(1), pp. 161-164.
Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. Computational linguistics, 29(3), pp. 333-347.
Siebenhaar, B. (2003). Sprachgeographische Aspekte der Morphologie und Verschriftung in schweizerdeutschen Chats. In: Linguistik Online, 15(3/03). Retrieved from http://www.linguistik-online.de/15_03/siebenhaar.html (March 9, 2016).
Siebenhaar, B. (2005). Die dialektale Verankerung regionaler Chats in der deutschsprachigen Schweiz. In: Eggers, E., Schmidt, J. E., & Stellmacher, D. (eds.): Moderne Dialekte – Neue Dialektologie, pp. 691-717. Stuttgart: Steiner.
Ziegler, E. (2005). Die Bedeutung von Interaktionsstatus und Interaktionsmodus für die Dialekt - Standard - Variation in der Chatkommunikation. In: Eggers, E., Schmidt, J. E., & Stellmacher, D. (eds.): Neue Dialekte – Moderne Dialektologie, p. 719-745. Stuttgart: Steiner.
Zehetner, Ludwig (1985). Das bairische Dialektbuch. München: Verlag C. H. Beck.
Zehetner, Ludwig (2014). Bairisches Deutsch. Lexikon der Deutschen Sprache in Altbayern. Regensburg: Edition Vulpes.