Bairisch 2.0 – Erstellung eines Social Media-Dialektlexikons mithilfe von Crowdsourcing

Manuel Burghardt

Abstract

This paper describes the creation of a Bavarian dialect lexicon that is based on social media and that was translated by means of crowdsourcing. In the process, we first obtain a corpus of dialect samples from the social network Facebook. For the translation of the dialect words to standard German, we involve the members of the social network, which allows us to produce a crowdsourced lexicon of Bavarian dialect words.

1 Dialektologie im digitalen Zeitalter

Unter einem Dialekt oder einer Mundart, versteht man gemeinhin eine regionale Sprachvarietät, die sich in Phonologie, Grammatik, Syntax und Wortschatz teilweise stark von der sog. Standardsprache unterscheidet. Wenngleich sich für einige Dialekte auch Beispiele der systematischen Verschriftlichung finden,[1] so ist Dialekt doch primär ein Phänomen gesprochener Sprache. Entsprechend aufwendig ist die Datenakquise in der Dialektologie[2], müssen doch erst konkrete Belege im Feld unter möglichst authentischen Bedingungen gesammelt werden. Nach dem Erheben des Dialektmaterials folgt zudem noch der arbeitsintensive Schritt der Transkription, um die Daten einheitlich zu dokumentieren. Demgegenüber steht die breite Verfügbarkeit von Sprachdaten in sozialen Netzwerken wie bspw. *Facebook*, in denen die Mitglieder häufig Dialekt sprechen, bzw. – im Sinne computervermittelter Kommunikation – schreiben. Die Vorteile dieser online verfügbaren Sprachdaten sind also (1) die enorme Verfügbarkeit, die quantitativ die Grenzen klassischer Dialektkorpora deutlich überschreitet, (2) die Authentizität der Sprecher*innen, die ganz ungezwungen Nachrichten im Dialekt formulieren, und (3) die erfreuliche Tatsache, dass

[1] So gibt es etwa für viele Wikipedia-Artikel eine bairische Übersetzung im Rahmen der „Boarischen Wikipedia“ siehe <https://bar.wikipedia.org/wiki/Hoamseitn>.

[2] Dialektologie bezeichnet den Teil der Sprachwissenschaft, der sich aus historischer, geographischer und sozialer Perspektive mit dem Phänomen Dialekt beschäftigt. Für eine umfassende Einführung siehe Löffler (2003).

die Kommunikation in schriftlicher Form erfolgt und somit bereits eine grundlegende Transkription vorliegt.

Freilich bringen diese Vorteile ganz eigene Herausforderungen und Probleme mit sich, bspw. wenn die Kommunikation über nicht-öffentliche Kanäle erfolgt und die Verfügbarkeit damit eingeschränkt ist. Wenngleich die sprachlichen Äußerungen in sozialen Netzwerken einerseits dahingehend authentisch sind, dass niemand die Sprecher*innen explizit zur Äußerung auffordert, so leidet die dialektale Authentizität doch andererseits durch den computervermittelten Übertragungskanal, bei dem sich ganz natürlich auch Merkmale von asynchroner Online-Kommunikation mit in die Äußerungen einschleichen (bspw. Emojis oder Chat-Sprache). Der zuletzt genannte Vorteil der automatischen Transkription birgt ebenfalls Gefahren hinsichtlich Objektivität und Konsistenz: Da die Sprecher*innen ihren Dialekt nicht einheitlich oder gar nach einem phonetischen Standard transkribieren, ist dementsprechend ein hohes Maß an orthografischer Varianz bei den online geschriebenen Dialektbegriffen zu erwarten. Aus Perspektive der Dialekt-Lexikologie fehlt schließlich eine entscheidende Information: die standardsprachliche Entsprechung bzw. Übersetzung eines Dialektworts, die nicht unmittelbar aus den sozialen Medien extrahiert werden kann.

In diesem Artikel soll demonstriert werden, dass die Vorteile der massenhaften Verfügbarkeit von geschriebenen Dialektdaten dabei die angesprochenen Nachteile und Defizite überwiegen. Konkret wird dabei ein Korpus für einen bairischen Dialekt (ISO 639-3:bar) automatisch aus dem sozialen Netzwerk Facebook gewonnen und anschließend standardsprachlich übersetzt. Für die Übersetzung binden wir die Mitglieder des sozialen Netzwerks Facebook ein und erzeugen so ein neuartiges Dialektlexikon auf Basis eines Crowdsourcing-Ansatzes (cf. Howe 2006). Der Beitrag reiht sich konzeptuell in eine noch junge Tradition von Studien ein, die das „Web als Korpus“ (Baroni et al. 2009; Kilgarriff / Grefenstette 2003) zur Untersuchung dialektologischer und soziolinguistischer Fragestellungen heranziehen (cf. z. B. Doyle 2014; Eisenstein 2015; Jørgensen et al. 2015; Siebenhaar 2005) und liefert erste Impulse für die Erstellung von Dialektlexika für den konkreten Anwendungsfall des Bairischen.[3]

[3] Zu diesem Thema wurde im Vorfeld bereits eine grundlegende Machbarkeitsstudie publiziert (cf. Burghardt et al. 2016). Weiterhin wurde im Rahmen der TEDx-Konferenz ein inhaltlich verwandter Vortrag zum Thema „Dialektsprache mit Hilfe

2 Bairischer Dialekt

Im deutschsprachigen Raum ist das Bairische ein weitverbreiteter Dialekt. An dieser Stelle sei ein orthografischer Hinweis zur Unterscheidung von „Bayrisch“ und „Bairisch“ erlaubt: Die y-Schreibung wurde vom hellenophilen König Ludwig I. eingeführt und bezeichnet das geografische Konstrukt, den Freistaat „Bayern“. Die i-Schreibung findet sich immer dann, wenn vom bairischen Dialekt die Rede ist. Diese Abgrenzung ist wichtig, da nicht alle Einwohner des Freistaats Bayern im bairischen Dialekt sprechen (sondern bspw. Schwäbisch, Thüringisch und Alemanisch) und gleichzeitig Bairisch auch an anderen Orten gesprochen wird, bspw. in Österreich (cf. Schmid 2012). Weiterhin ist der bairische Dialekt ein stark heterogenes Sprachkonstrukt, dass sich nach Zehetner (1985) in mindestens fünf größere Dialektfamilien einteilen lässt (siehe Tab. 1).

Dialektfamilie	Regionen
Nordbairisch	Nördliche Oberpfalz / östliches Oberfranken, westliche Oberpfalz / östliches Mittelfranken, Mittlere Oberpfalz
Nordmittelbairisch	Südliche Oberpfalz / nördliches Niederbayern, mittlerer Bayerischer Wald
Mittelbairisch	Unterer Bayerischer Wald, Ober- und Niederbayern, westliches Oberbayern
Südmittelbairisch	Oberbayerisches Alpengebiet
Südbairisch	Werdenfelser Land, Isarwinkel

Tab. 1: Überblick zu bairischen Dialektfamilien und den Regionen, in denen diese hauptsächlich gesprochen werden

Aufgrund der persönlichen Dialektkompetenz und der Tatsache, dass das Mittelbairische die Dialektfamilie mit den meisten Sprecher*innen darstellt (cf. Zehetner 1985), wird in der nachfolgend beschriebenen Fallstudie vorrangig diese Dialektfamilie untersucht.

---

von Facebook erforschen“ gehalten, der bei youtube verfügbar ist (cf. Burghardt 2017). Mein herzlicher Dank gebührt Daniel Granvogl, der im Rahmen seiner Masterarbeit im Fach Medieninformatik maßgeblich zur Umsetzung der hier beschriebenen Studie beigetragen hat.

3 Korpuserstellung

Auf dem sozialen Netzwerk Facebook gibt es für Mitglieder die Möglichkeit, sich in thematischen Gruppen zu organisieren. Solche Gruppen können entweder für einen geschlossenen Kreis (bspw. für die Organisation einer privaten Feierlichkeit) oder aber für die breite Öffentlichkeit erstellt werden, wobei letzteren Gruppen jeder beitreten kann. Für unsere Zwecke sind vor allem solche öffentlichen Gruppen relevant, die sich explizit dem Sprechen (bzw. Schreiben) von bairischem Dialekt verschrieben haben. Ein wesentliches Auswahlkriterium bei der Vielzahl bestehender Dialekt-Gruppen war eine bestimmte Mindestgröße (mind. 500 Mitglieder), Aktivität (es werden regelmäßig Nachrichten in der Gruppe veröffentlich) sowie eine primäre Fokussierung auf den mittelbairischen Dialekt:

- *Boarisch Bluad*: 14.733 Mitglieder[4]
  Quelle: <https://www.facebook.com/groups/248236578706001>
- *Boarisch redn is in*: 5.071 Mitglieder
  Quelle: <https://www.facebook.com/groups/328341027261813/>
- *Niederboarisch für Anfänger und Runaways*: 766 Mitglieder
  Quelle: <https://www.facebook.com/groups/121572707986445/>

Für die Erstellung des Dialektkorpus wurden zunächst sämtliche Nachrichten (siehe Abb. 1) aus den oben genannten Gruppen über die *Facebook Graph API*[5] heruntergeladen und im *JSON*-Format (*JavaScript Object Notation*) gespeichert.

[4] Alle Mitgliederzahlen beziehen sich auf den Zeitraum der Korpuserstellung, die bereits im November 2016 erfolgte.

[5] Zur *Facebook Graph API* siehe <https://developers.facebook.com/docs/graph-api>.

Abb. 1: Beispielhafte Dialekt-Nachricht aus einer öffentlichen Facebook-Gruppe, die sich bei der Gruppenkommunikation primär dem Dialekt verschrieben hat

Abbildung 2 zeigt die formal-technische JSON-Repräsentation einer beispielhaften Nachricht in einer der Gruppen. Wir speichern dabei die URL zur entsprechenden Nachricht, Name (hier teilanonymisiert) und ID des Autors / der Autorin, das Veröffentlichungsdatum, den Nachrichtentext sowie die Anzahl der „Gefällt mir"-Angaben (*Likes*).

dialect.json

```
{
   "url" : "https://www.facebook.com/groups/121572707986445/permalink/567546253389086/",
   "user" : {
       "name" : "XXXX Andrea",
       "id" : "1031233483572672"
   },
   "date" : "2015-01-08T08:32:04+0000",
   "message" : "Grisdeich, Boa i  war grod beim Beck und doa is mia da giggal gsting,
   weggal hoisn bei uns jez Schrippen. I glaubs ned.",
   "likes_count" : NumberInt(3),
}
```

Abb. 2: JSON-Repräsentation der Nachricht.

Im Ergebnis erhält man so 213.292 Nachrichten von 9.164 unterschiedlichen Mitgliedern. Als Nächstes erfolgt die Erstellung einer Frequenzliste für die insgesamt 101.056 unterschiedliche laufende Wortformen[6] ermittelt wurden. Offenkundig sind nicht alle Wörter in den Nachrichten Dialektwörter, denn teilweise findet sich eine Mischung aus Dialekt- und Standardsprache. Noch häufiger allerdings findet sich eine Mischung aus Dialekt- und Chat-Sprache (*lol*, *omg*, etc.). Ein wichtiger Schritt ist somit das Herausfiltern der bairischen Begriffe aus den Sprachdaten.

In einer vorangegangenen Studie (cf. Burghardt et al. 2016) wurden nicht-bairische Begriffe automatisiert entfernt, indem ein Korpus nicht-di-

[6] Auf die in der automatischen Sprachverarbeitung sonst übliche Grundformenreduktion der Sprachdaten muss hier verzichtet werden, da es (noch) keine automatischen Lemmatisierungsverfahren für bairische Dialektwörter gibt.

alektaler Onlinesprache (cf. Beißwenger 2013) als erweiterte Stoppwortliste verwendet wurde. Leider gingen auf diese Weise viele Dialektwörter verloren, die Homographen zu standardsprachlichen Wörtern sind, beispielsweise bairisch *nixe* (‚nichts') und standardsprachlich *Nixe* (‚Meerjungfrau') sowie bairisch *affe* (‚hinauf') und standardsprachlich *Affe* (‚Primat').[7] Für die hier beschriebene Folgestudie wurde deshalb ein qualitatives Auswahlverfahren gewählt, d. h. aus der zuvor erstellten Frequenzliste wurden per Hand Dialektwörter aus den 500 häufigsten Wörtern extrahiert.[8] Im Ergebnis bleiben 259 hochfrequente Dialektwörter übrig. Tabelle 2 zeigt die 25 häufigsten Dialekt-Wörter (s. u. Tab. 2). Dabei wird deutlich, dass es sich hierbei primär um phonetische Varianten hochfrequenter standardsprachlicher Wörter handelt (*auch*, *ich*, *ist*, ...). Wörter, die aus lexikalischer Perspektive interessant sind, erscheinen erst im Mittelfeld der Frequenzliste, bspw. *zefix* (Schimpfwort im Sinne von ‚verdammt nochmal'; Häufigkeit 729) oder *bua* (‚Junge, Bub'; Häufigkeit 320).

Zu den entfernten Nicht-Dialektwörtern zählen in erster Linie Eigennamen (vor allem Namen der Gruppenmitglieder), aber auch häufige Wörter der Standard- und Chat-Sprache (bspw. *ja*, *mehr*, *Tag*, *lach*, etc.). Für die 259 Dialektwörter wurde eine Konkordanz mit bis zu zehn Belegsätzen erstellt und in einer Datenbank gespeichert. Nachfolgend findet sich ein Überblick zu beispielhaften Dialektsätze für den bairischen Partikel *fei*.[9]

(1) *ja Gaby des hama mir als Kinda imma gmachd :) und dann af d´Nachd s´Lichd ausdrim na hod ma des fei wirklich gsegn ....*

(2) *A packal Watschn is fei glei afgrissn*

(3) *des hod fei zu mia heid Nachd oana gsagd: er dringd de Noagal scho zsam - do san teilweis fasd voie Glasl umanand gschdana weis voa lauda Rausch nix mea oibrod ham*

(4) *oiso i difdld fei scho hin und wieder an gscheidn Bleedsin aus* ;-)

---

[7] Für die automatische Korpusverarbeitung wurden sämtliche Begriffe in Kleinbuchstaben umgewandelt, sodass die Groß- und Kleinschreibung kein hinreichendes Unterscheidungskriterium darstellt.

[8] Dieses Vorgehen ist auch dadurch motiviert, dass wir in der nachgeschalteten Übersetzung mittels Crowdsourcing ohnehin nur eine begrenzte Anzahl an Dialektwörtern übersetzen lassen können.

[9] Für eine Diskussion des Wortes siehe Schmid (2012).

(5) *dua fei koa Staffe überseng sonst foist hi*

(6) *Des hob i fei friara scho a moi gmacht bei meim Onkel aufm Hof.*

(7) *Do hertse fei da spass af gei*

(8) *schlaf fei ned undam geh ei*

(9) *werds fei ned scho wida grob - an Blazaldoag dama damid auswoigln sisd nix!*

(10) *hmmm - kenn i fei a ned. Petra des is a anders Gai.*

	Bairischer Begriff	**Übersetzung**	**Häufigkeit im Korpus**
1	*a*	auch	41066
2	*i*	ich	36642
3	*is*	ist	21599
4	*guad*	gut	15917
5	*ned*	nicht	14191
6	*de*	die	13862
7	*moang*	morgen	10737
8	*scho*	schon	9693
9	*ma*	wir	9178
10	*do*	da	8366
11	*no*	noch	8336
12	*moing*	morgen	7090
13	*mei*	man	6251
14	*wos*	was	6177
15	*eich*	euch	5746
16	*hob*	ich habe	5248
17	*scheena*	schöner	4883
18	*mi*	mich	4129
19	*nix*	nichts	3447
20	*hod*	hat	3394
21	*schee*	schön	3074
22	*wia*	wie	3050

23	*dog*	tag	3023
24	*moi*	mal	2934
25	*ois*	alles	2855

Tab. 2: Überblick zu den 25 häufigsten Dialektwörtern, deren standardsprachlicher Übersetzung sowie deren Häufigkeit im Gesamtkorpus.

4 Übersetzung mithilfe von Crowdsourcing

Den nächsten Schritt bei der Erstellung eines Dialektlexikons des Mittelbairischen stellt die Übersetzung der Dialektwörter mithilfe eines Crowdsourcing-Ansatzes dar. Das Kofferwort Crowdsourcing wurde maßgeblich von Jeff Howe (2006) geprägt und beschreibt im Wesentlichen das Auslagern (*outsourcing*) von Aufgaben an Personen, die nicht im herkömmlichen Sinne Experten mit einer expliziten Qualifikation für die Bewältigung der Aufgabe sind (*crowd*). Entsprechend müssen die Aufgaben so gewählt und präsentiert werden, dass auch Laien ohne Vorkenntnisse sie erledigen können. Typische Aufgaben im Bereich Digital Humanities und digitales Kulturerbe sind dabei die Klassifikation von Objekten (bspw. Texte oder Bilder), die Erweiterung von digitalen Sammlungen (bspw. Fotos oder Sprachaufnahmen) oder die Korrektur und Transkription von Texten (cf. Oomen / Aroyo 2011). Für den zuletzt genannten Aufgabentyp finden sich besonders viele Beispiele, bspw. das *Transcribe Bentham*-Projekt zur Transkription der handschriftlichen Transkripte von Jeremy Bentham (cf. Causer / Wallace 2012) oder das Allegro-Projekt zur Transkription von handschriftlichen Melodien historischer Volkslieder (cf. Burghardt / Spanner 2017). Transkription lässt sich auch deshalb gut in die *crowd* auslagern, weil die Aufgabe – wenn nicht gerade komplexe Transkriptionsrichtlinien berücksichtigt werden müssen – relativ intuitiv bewerkstelligt werden kann und kein weiteres Vorwissen benötigt wird. Auch die Umsetzung der Aufgabenstellung im Sinne eines Crowdsourcing-Tools[10] ist verhältnismäßig einfach, da im Wesentlichen nur die zu übersetzende Quelle dargestellt werden muss und ein Eingabefeld für die Transkription vorhanden sein muss.

[10] Crowdsourcing impliziert hier automatisch die Nutzung des Webs als Plattform und die Umsetzung der Aufgaben mithilfe digitaler Tools und Webservices. Digitalität ist dabei aber kein hartes Kriterium für Crowdsourcing. So finden sich etwa frühe Beispiele für Crowdsourcing in der Ornithologie, die gänzlich ohne digitale Netzwerke und Tools auskommen. Richtig ist aber, dass mit der Verfügbarkeit des

Auch im vorliegenden Fall, der Erstellung eines Dialektlexikons auf Basis von online verfügbaren Dialektsätzen, liegt eine Transkriptionsaufgabe vor, bei der Dialektwörter in Standardsprache übersetzt werden sollen. Eine zentrale Herausforderung bei Crowdsourcing-Ansätzen liegt in der Akquise entsprechender *crowdworker*. Sofern keine finanziellen Anreize für das Erledigen einer Aufgabe geschaffen werden können, gilt es, entsprechende Personen auf anderem Wege zu motivieren. Dies gelingt mitunter durch eine ansprechende, benutzerfreundliche Präsentation der Aufgabe (Burghardt / Spanner 2017), die im Idealfall sogar spielerisch im Sinne von Gamification (cf. Deterding et al. 2011) umgesetzt werden kann. Zentral für das Gelingen eines Crowdsourcing-Ansatzes ist allerdings, dass sich für die Erledigung der Aufgabe ein intrinsisch motivierter Nutzerkreis findet, der auf freiwilliger Basis ein Interesse an der Erledigung der Aufgabe hat.[11] Die Bedingungen für erfolgreiches Crowdsourcing in der vorliegenden Studie sind außerordentlich günstig, da die Mitglieder expliziter Dialekt-Gruppen ein hohes Maß an intrinsischer Motivation mit sich bringen, wenn es darum geht ihre eigene Dialektsprache zu übersetzen: Da Dialekt auf der Bedeutungsebene hochgradig ambig ist, besteht ein gesteigertes Interesse an einem spezifischen Dialektlexikon, in dem nicht nur die eigene Übersetzung, sondern auch potenzielle Varianten anderer enthalten sind.

Für die Übersetzungsaufgabe wurde eine Webanwendung erstellt (siehe Abb. 3, links), die sowohl an Bildschirmarbeitsplätzen als auch auf mobilen Geräten durch das einfache Klicken auf einen Hyperlink ausgeführt werden kann. Beim Design des Interface wurde durch vorherige Benutzertests das Augenmerk auf ein hohes Maß an Intuitivität und Benutzerfreundlichkeit gelegt, da möglichst viele der Gruppenmitglieder am Übersetzungsprozess teilnehmen sollten. Darüber hinaus wurden unterschiedliche Gamification-Aspekte implementiert, bspw. ein einfaches System zur Kommunikation des eigenen Fortschritts sowie auch ein interaktiver Avatar, der die Anwender*innen durch die Aufgabenstellung des

Internets auch das Crowdsourcing in ganz anderen Dimensionen möglich ist, da im Prinzip jeder mit Zugang zu einem webfähigen Computer Teil einer virtuellen *crowd* werden kann (siehe etwa Amazons Plattform *Mechanical Turk*: <https://www.mturk.com>).

[11] Siehe Oomen und Aroyo (2011) für eine ausführlichere Diskussion der Motivation von Teilnehmer*innen an einem Crowdsourcing-Projekt.

Übersetzens führt und sie durch aufmunternde Worte in Dialektsprache immer wieder zum Weitermachen ermuntert (siehe Abb. 3, rechts).

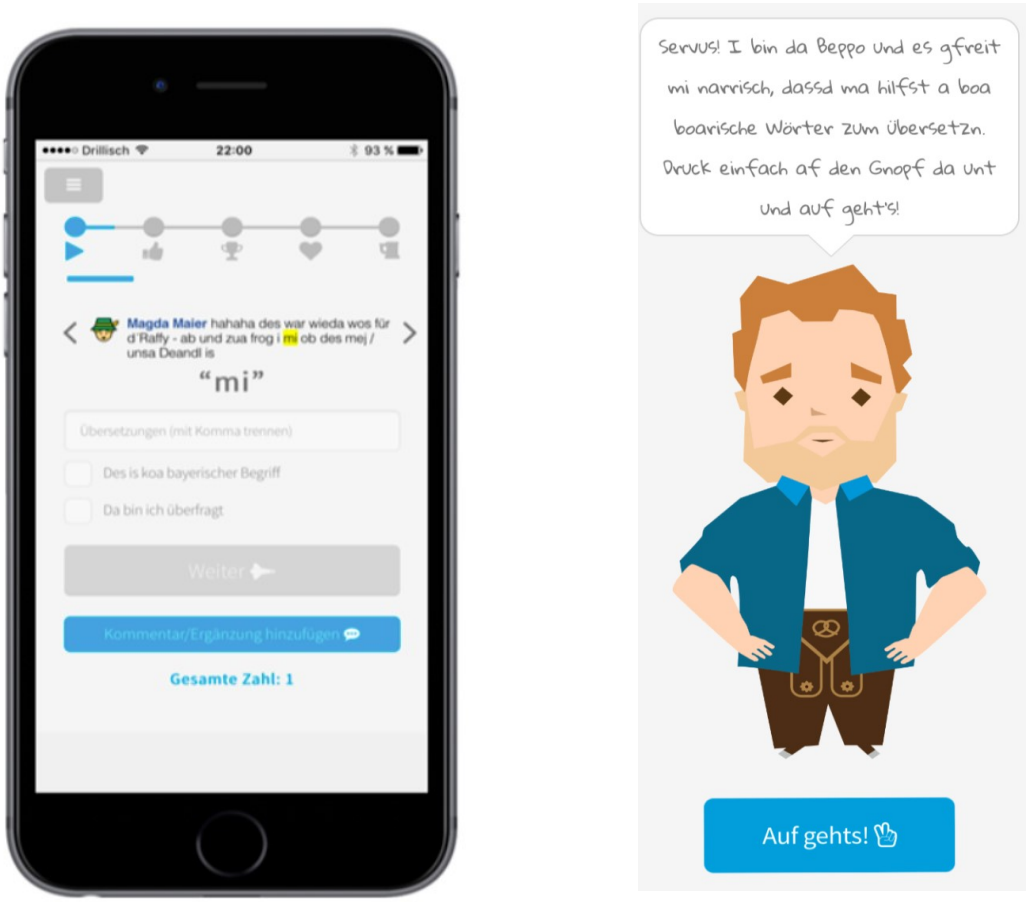

Abb. 4: Links: Darstellung der Übersetzungsaufgabe auf einem mobilen Endgerät. Rechts: Darstellung des Avatars „Beppo", der die Übersetzer*innen während der Aufgabe begleitet und sie immer wieder ermuntert und motiviert die Aufgabe zu vollenden.

Die konkrete Aufgabenpräsentation zeigt einen dialektalen Belegsatz aus dem Facebook-Korpus an, in dem das gesuchte Dialektwort hervorgehoben ist. Darunter findet sich ein Freitextfeld für die Übersetzung. Weiterhin gibt es die Option „Des is koa bairischer Begriff" (es handelt sich nicht um ein Dialektwort) und „Da bin ich überfragt" (das Wort ist nicht bekannt und / oder kann nicht in die Standardsprache übersetzt werden) sowie die Möglichkeit eigene Kommentare hinzuzufügen. Die Kommentarfunktion wurde vor allem bei stark ambigen Wörtern (bspw. *fei*) verwendet. In einer Vorstudie wurde deutlich, dass Teilnehmer*innen im Schnitt ca. 25 Wörter übersetzen, bevor sie die Aufgabe abbrechen (cf. Burghardt et al. 2016). Aus diesem Grund liegt die Obergrenze der zu übersetzenden Wörter pro Teilnehmer*in in dieser Studie bei 25 Begriffen. Zu

Beginn werden die 259 Dialektwörter den Teilnehmer*innen nach dem Zufallsprinzip zugewiesen. Nachdem die ersten Übersetzungen vorliegen, werden dann systematisch diejenigen Wörter zum Übersetzen angezeigt, die bislang am wenigsten Übersetzungen aufweisen. So wird sichergestellt, dass für die unterschiedlichen Worte annähend gleich viele Übersetzungen vorliegen.

Die Crowdsourcing-Übersetzungsaufgabe dauerte insgesamt 4 Wochen an. Innerhalb dieser Zeit wurden 3.387 Übersetzungen von 291 Teilnehmer*innen erstellt, das bedeutet im Schnitt etwa 13 Übersetzungen für jeden der 259 häufigsten Dialekt-Begriffe im Facebook-Korpus. Die Ergebnisse wurden den Mitgliedern der Facebook-Gruppen als Belohnung für ihre Teilnahme an der Crowdsourcing-Studie in Form eines einfachen Dialektlexikons zur Verfügung gestellt. Sämtliche Daten liegen auch strukturiert vor und können für weitere dialektologische Untersuchungen genutzt werden.[12]

5 Diskussion

In diesem Beitrag wurde ein neuartiger Ansatz zur Erstellung eines Dialektlexikons des Bairischen auf Basis von Social Media-Sprachdaten vorgestellt, der mithilfe eines Crowdsourcing-Ansatzes von der Community selbst übersetzt wurde. Mit dieser Studie soll das Potenzial von online verfügbaren Dialektdaten für dialektologisch-lexikographische Fragestellungen aufgezeigt werden, dass bestehende Langzeitprojekte, wie etwa das Bayerische Wörterbuch (BWB)[13], natürlich nicht ersetzen kann, wohl aber weiterführende Perspektiven für die Erstellung von Dialektlexika auf Basis aktueller Sprachbelege aufzeigen soll.

Die Neuerungen des hier präsentierten Verfahrens bestehen einerseits in der Datenakquise und Korpuserstellung und andererseits in der Übersetzung der Daten mittels Crowdsourcing. Als wesentlicher Vorteil der Nutzung von Social Media-Daten muss deren Aktualität hervorgehoben werden. Klassische Wörterbuchprojekte wie das BWB stützten sich vornehmlich auf (mittlerweile) historische Belege, die seit 1913 fortlaufend erhoben wurden. Wenngleich diese diachrone Perspektive unverzichtbar

[12] Die Daten sind online als Tabelle verfügbar unter <https://bit.ly/2RLDWdA>.

[13] Akademieprojekt Bayerisches Wörterbuch, siehe <https://www.bwb.badw.de/bairische-mundarten.html>.

für ein umfassendes Lexikon ist, so scheint die systematische Dokumentation aktueller Belege ebenso wichtig. Ein weiterer Vorteil liegt in der Authentizität der Facebook-Daten, da die Sprachbelege freiwillig und ungezwungen von den Sprecher*innen erstellt werden. Diese Ungezwungenheit bringt gleichzeitig auch einige Nachteile mit sich: So fehlen für die meisten Belege systematische Metadaten, wie bspw. Geschlecht, Alter, Herkunftsort, etc. Weiterhin bestehen gewisse Verzerrungen durch die Verschriftlichung des Dialekts, die sicherlich nur näherungsweise die tatsächliche Aussprache widerspiegelt. Wie oben beschrieben, ergeben sich weitere Verzerrungen durch den Kommunikationskanal Internet, bei dem häufig dialektale Sprachbelege mit solchen typischer Onlinekommunikation vermischt werden.

In Hinblick auf die Übersetzung mittels Crowdsourcing soll als positiver Aspekt nochmals die Verfügbarkeit hochgradig motivierter Übersetzer*innen betont werden. Ein Blick auf das so erstellte Dialektlexikon (siehe Tab. 3) offenbart zugleich aber auch die eingangs erwähnten Probleme eines solchen Ansatzes, der häufig Inkonsistenzen und in einzelnen Fällen auch Fehler aufweist. So zeigt sich etwa bei den Übersetzungen des Dialektworts *voi*, das neben der korrekten Übersetzung ‚voll' auch fälschlicherweise ‚viel' (aus den Belegsätzen ergibt sich diese Bedeutung nicht) übersetzt wird. Offenbar empfinden einige der Laien-Übersetzer den Verstärkungspartikel *voll* als zu umgangssprachlich und paraphrasieren ihn deshalb mit (vermeintlich) standardsprachlichen Begriffen (aus ***voll*** *gut* wird dann bspw. ***total*** *gut* oder ***richtig*** *gut*). Einige umschreiben den Begriff bzw. dessen Bedeutung durch Adjektive wie *verstärkt*, *intensiv*, *extrem*. All diese Varianten sind offenkundig das Resultat fehlender Übersetzungsrichtlinien und objektivierbaren Regeln zur einheitlichen Übersetzung. Positiv hervorzuheben ist hier, dass die Mehrheitsübersetzung des Begriffs (9 von 16) der korrekten Übersetzung ‚voll' entspricht. Weiterhin ist zu beobachten, dass Nutzer teilweise die angeführten Belegsätze komplett zu ignorieren scheinen und das jeweilige Wort isoliert übersetzen. Dies führt häufig zu Inkonsistenzen, bspw. wenn der mehrdeutige Begriff *wei* nicht – wie durch die jeweiligen Belegsätze vorgegeben – als ‚weil', sondern als ‚Weib' oder paraphrasiert als ‚Frau' übersetzt wird. Auch hier findet sich wieder ein Beispiel einer sinngemäßen Paraphrase, indem *weil* als ‚denn' übersetzt wird. Auch hier kann man allerdings ins Feld führen, dass die Mehrheitsübersetzung korrekt ist. Eine weitere Inkonsistenz wird beim Begriff *ghabt* deutlich. Wie eingangs beschrieben,

lassen wir laufende Wortformen übersetzen. Die Mehrheit der Übersetzer*innen führt hier korrekt die standardsprachliche Übersetzung ‚gehabt' an, eine Person gibt als Übersetzung allerdings das Lemma ‚haben' an. Es ist an dieser Stelle festzuhalten, dass der Crowdsourcing-Ansatz – wie erwartet – tatsächlich viele Inkonsistenzen und teilweise auch Fehler beinhaltet. Gleichzeitig ist das Mehrheitsvotum in den meisten Fällen aber korrekt. Bei der Verwendung der Daten sollte also vor allem die mehrheitlich angegebene Übersetzung berücksichtigt werden. Alle weiteren Varianten sollten kritisch geprüft werden um herauszufinden, ob hier wirklich eine linguistisch-relevante Mehrdeutigkeit besteht (wie etwa beim nicht eindeutig zu übersetzenden Partikel *fei*) oder ob die Varianz ggf. lediglich aus Unklarheiten und Missverständnissen bei der Übersetzungsaufgabe resultiert. Für künftige Studien nach diesem Ansatz empfehlen wir dringend den *crowdworkern* grundlegende Übersetzungsregeln und Beispielübersetzungen zur Verfügung zu stellen. Idealerweise erfolgt vor der eigentlichen Übersetzungsaufgabe eine kurze Einführung in Form eines interaktiven Tutorials.

Dialektwort	Korpushäufigkeit	Übersetzungen	Häufigkeit der jeweiligen Übersetzung
voi	451	voll	9
		viel	2
		verstärkt	1
		total	1
		richtig	1
		intensiv	1
		extrem	1
wei	437	weil	12
		denn	1
		weib	1
		frau	1
ghabt	458	gehabt	12
		haben	1
		getragen	1

fei	436	aber	8
		user_didnt_know	3
		im Übrigen	1
		gell	1
		sowieso	1

Tab. 3: Auszug aus dem mittels Crowdsourcing erstellten Dialektlexion.

Die eingangs beschriebenen Inkonsistenzen bei der Transkription, die keinen einheitlichen Richtlinien folgt, offenbart zunächst einen erheblichen Nachteil. Andererseits kann durch einen demokratischen Übersetzungsansatz argumentiert werden, dass keine subjektive Verzerrung durch wenige Experten-Redakteure erfolgt, sondern vielmehr ein Mehrheitsvotum aus vielen parallelen Übersetzungen gezogen werden kann. Nachteilig ist hier wiederum das bereits angesprochene Fehlen von Metadaten, die es bspw. erlauben würden Unterschiede in der Übersetzung mit regionalen oder altersspezifischen Variablen in Verbindung zu setzen.

6 **Abschließende Betrachtungen**

Die hier präsentierte Studie belegt das Potenzial nutzergenerierter Dialektbelege und entsprechender Übersetzungen, zeigt aber gleichzeitig Einschränkungen im Bereich der Korpuserstellung und des Crowdsourcing auf, die in Folgestudien systematisch adressiert werden müssen. Für eine Übertragbarkeit des hier beschriebenen Ansatzes auf andere Dialekte ist hervorzuheben, dass sich auf Facebook auch viele weitere Dialekt-Gruppen jenseits des Bairischen finden, bspw. die öffentliche Gruppe „Plattdeutsch – the language of champions!“, die mehrere tausend Mitglieder aufweist.[14] Eine Erweiterung bestehender dialektologischer Arbeitspraxen um den in diesem Artikel vorgestellten Ansatz eröffnet neuartige Forschungsperspektiven, die gleich in zweifacher Hinsicht Charakteristika der *Digital Humanities*[15] aufgreifen: Einerseits kommen *digitale Tools* im Sinne der computergestützten Korpuserstellung zum Einsatz, andererseits

[14] Facebook-Gruppe „Plattdeutsch“: <https://www.facebook.com/groups/8313036843/>.

[15] Siehe hierzu auch die Taxonomie von Roth (2019).

wird durch die Analyse von Social Media-Daten ein genuin *digitales Kulturphänomen* untersucht.

Bibliographie

Baroni, Marco / Bernardini, Silvia / Ferraresi, Adriano / Zanchetta, Eros (2009): „The wacky wide web: a collection of very large linguistically processed web-crawled corpora", in: *Language Resources and Evaluation* 43, 3: 209-226.

Beißwenger, Martin (2013): „Das Dortmunder Chat-Korpus", in: *Zeitschrift für germanistische Linguistik* 41, 1: 161-164.

Burghardt, Manuel (2017): „Dialektsprache mit Hilfe von Facebook erforschen". Vortrag im Rahmen von TEDxOTHRegensburg <https://www.youtube.com/watch?v=ciWFd-1O4KU> [29.09.2021].

Burghardt, Manuel, Granvogl, Daniel / Wolff, Christian (2016): „Creating a Lexicon of Bavarian Dialect by Means of Facebook Language Data and Crowdsourcing", in: Maegaard, Bente / Calzolari, Nicoletta / Choukri, Khalid (eds.): *Proceedings of the 10th edition of the Language Resources and Evaluation Conference* (LREC). Paris: European Language Resources Association 2029-2033.

Burghardt, Manuel / Spanner, Sebastian (2017): „Allegro: User-centered Design of a Tool for the Crowdsourced Transcription of Handwritten Music Scores", in: *Proceedings of the DATeCH (Digital Access to Textual Cultural Heritage) conference*. Association for Computing Machinery 15-20.

Causer, Tim / Wallace, Valerie (2012): „Building A Volunteer Community: Results and Findings from Transcribe Bentham" in: *Digital Humanities Quarterly* 6, 2.

Deterding, Sebastian / Sicart, Miguel / Nacke, Lennart / O'Hara, Kenton / Dixon, Dan (2011): „Gamification. Using game-design elements in non-gaming contexts", in: *CHI'11. Extended abstracts on human factors in computing systems*. Association for Computing Machinery 2425-2428.

Doyle, Gabriel (2014): „Mapping dialectal variation by querying social media", in: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics 98-106.

Eisenstein, Jacob (2015): „Systematic patterning in phonologically-motivated orthographic variation", in: *Journal of Sociolinguistics* 19, 2): 161-188.

Howe, Jeff (2006): „The Rise of Crowdsourcing", in: *Wired* <https://www.wired.com/2006/06/crowds/> [06.10.2021].

Jørgensen, Anna Katrine / Hovy, Dirk / Søgaard, Anders (2015): „Challenges of studying and processing dialects in social media", in: *Proceedings of the Workshop on Noisy User-generated Text*. Beijing: Association for Computational Linguistics 9-18.

Kilgarriff, Adam / Grefenstette, Gregory (2003): „Introduction to the special issue on the web as corpus", in: *Computational Linguistics* 29, 3: 333-347.

Löffler, Heinrich (2003): *Dialektologie*. Eine Einführung. Tübingen: Narr.

Oomen, Johan/ Aroyo, Lora (2011): „Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges", in: *C&T '11: Proceedings of the 5th International Conference on Communities and Technologies*. New York: Association for Computing Machinery 138–149.

Roth, Camille (2019): „Digital, digitized, and numerical humanities", in: *Digital Scholarship in the Humanities* 34, 3: 616-632.

Schmid, Hans Ulrich (2012): *Bairisch – Das Wichtigste in Kürze*. München: Verlag C. H. Beck.

Siebenhaar, Beat (2005): „Die dialektale Verankerung regionaler Chats in der deutschsprachigen Schweiz", in: Eggers, Eckhard / Schmidt, Jürgen E. / Stellmacher, Dieter (eds.): *Moderne Dialekte – Neue Dialektologie*. Stuttgart: Steiner 691-717.

Zehetner, Ludwig (1985): *Das bairische Dialektbuch*. München: Verlag C. H. Beck.