ALEKSANDER WIATR

BEDEUTUNG UND FUNKTION VON CROWDSOURCING IM PROJEKT VERBAALPINA

COBISS: 1.02

Pomen in vloga množičnega zunanjega izvajanja v projektu VerbaAlpina

Prispevek daje pregled projekta VerbaAlpina. Ta si prizadeva za sistematično analizo alpskega besedja, ki prehaja jezikovne meje, v zvezi z realnostjo, ki se tiče Alp. Prvi del je posvečen metodološkim osnovam. Pri tem so nakazani izzivi, ki so nastopili pri obdelavi zgodovinskih podatkov. Pojasnjeno je, kako so ti podatki preslikani v strukturirano in primerljivo obliko v podatkovni zbirki. Zadnji del se ukvarja z metodo množičnega zunanjega izvajanja (crowdsourcing), s katero se odpravijo obstoječe vrzeli in nekonsistentnosti in pridobi novo jezikovno gradivo.

Ključne besede: VerbaAlpina, množično zunanje izvajanje, medjezikovna geolingvistika, stratigrafija

The Meaning and Function of Crowdsourcing in the VerbaAlpina Project

This article offers an overview of the VerbaAlpina project. This is an effort to systematically analyze Alpine vocabulary that crosses linguistic boundaries in connection with the realities of the Alpine area. The first part is dedicated to methodological bases and presents the challenges that arose in processing historical data. It explains how the data were mapped into a structured and comparable form in the database. The last part deals with the methodology of crowdsourcing, whereby existing gaps are filled and inconsistencies corrected, and new linguistic material is collected.

Keywords: VerbaAlpina, crowdsourcing, inter-language geolinguistics, stratigraphy

1 VERBAALPINA: FORSCHUNGSGEGENSTAND UND ZIELE

VerbaAlpina ist ein durch die Deutsche Forschungsgemeinschaft gefördertes Projekt, das an der Ludwig-Maximilians-Universität München durchgeführt wird. In der ersten Laufphase wurde eine Finanzierung zunächst für drei Jahre (von Oktober 2014 bis Oktober 2017) – mit Aussicht auf Verlängerung bis Oktober 2025 – bewilligt. Das Projekt ist der Untersuchung des alpinen Raums, der durch die Grenzen der Alpenkonvention (http://www.alpconv.org/) bestimmt ist, gewidmet. Dieser umfasst 6989 politische Gemeinden aus der Schweiz, aus Italien, Frankreich, Österreich, Lichtenstein, Slowenien und Deutschland.[1]

1 Eine visuelle Darstellung der geographischen Grenzen des Untersuchungsgebietes bietet die projekteigene interaktive Karte. Abrufbar unter: https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=133 [letzter Zugriff: 14. 12. 2016].

Die Auswahl dieses spezifischen Gebiets ergibt sich aus der besonderen Stellung des Raums hinsichtlich seiner kulturellen und sprachlichen Geschichte. Aus ethnographischer und topographischer Perspektive können wir tendenziell von Homogenität sprechen, während das Gebiet aus sprachlicher Sicht von auffälliger Heterogenität geprägt ist (vgl. Krefeld – Lücke 2014: 189).

Seit der vorrömischen Zeit stellen die Alpen einen Raum dar, in dem das Aufeinandertreffen ganz unterschiedlicher Sprachgemeinschaften stets mit starken Akkulturationsprozessen einherging (vgl. Krefeld 2015d). Ab dem 6. Jh. n. Ch. wird das Gebiet durch eine Sprachkontaktsituation zwischen drei großen Sprachfamilien geprägt: der romanischen, der germanischen und der slawischen (vgl. Minnich 1989: 163). Tiefere Blicke in die Vergangenheit und in das Sprachmaterial verdeutlichen, wie vielfältig und kompliziert die sprachliche Stratifizierung der Alpen ist. Die heutigen Sprachen und Dialekte wie beispielsweise Französisch, Deutsch, Italienisch, Slowenisch, Friaulisch, Ladinisch, Rätoromanisch usw. existieren in diesem Raum nebeneinander. Kulturelle Gemeinsamkeiten des alpinen Lebens spiegeln sich in den Sprachen wider (vgl. Krefeld – Lücke 2014: 190–194) und werden im Projekt analytisch erschlossen. Im Fokus von VerbaAlpina stehen die alpine Lexik (nach Hubschmid 1951: 7) und alpentypische Realia, wie Almwirtschaft, Pflanzenwelt und der moderne Wintersport.

VerbaAlpina arbeitet dabei mit einer einzelsprachübergreifenden virtuellen Forschungsumgebung, die die folgenden Funktionsbereiche umfasst:

- *Dokumentation:* Die Dokumentation mittels interaktiver georeferenzierter Kartographie dient dazu, die bereits existierenden Sprachatlanten, dialektalen Wörterbücher, aber auch digitalen Projekte aus dem Untersuchungsgebiet zusammenzuführen und in Form einer Open-Source-Publikation (vgl. Krefeld 2015c) zu veröffentlichen, so dass eine grenzüberschreitende Perspektive geboten werden kann.
- *Datenerhebung und -ausgleich:* Um die existierenden Inkonsistenzen auszugleichen, werden mithilfe von Social Software (s. Kapitel 3 unten) neue Daten erhoben, alte Daten ergänzt und der Datenbestand validiert. Innovativ ist dabei der Zugriff auf das *Crowd* und die Tools des Crowdsourcings (s. Punkt 3.4).
- *Publikation und eigene kollaborative Weiterentwicklung:* Das Projekt ist als virtuelle Forschungsumgebung für Wissenschaftler/-innen und Laien gedacht, in deren Rahmen geforscht und publiziert werden kann. VerbaAlpina sieht sich also als Austauschplattform für zwei völlig unterschiedliche Adressatengruppen. Dieser Austausch soll dabei maßgeblich zur weiteren Erforschung alpiner Kulturen und Sprachen beitragen.

Im Gegensatz zur Vorgehensweise der klassischen Dialektologie verfolgt das Projekt nicht die Absicht, Dialekte als spezifische Varietäten herauszuarbeiten und voneinander abzugrenzen, sondern im Gegenteil das Netzwerk der gemeinsamen

Varianten zwischen diesen dialektalen Varietäten in einer grenz- und sprachübergreifenden Perspektive im Sinne einer interlingualen Sprachgeographie (vgl. Krefeld 2015d) darzustellen.

2 DIGITALISIERUNG UND AUFBEREITUNG DER HISTORISCHEN DATEN

Die Grundlage der in VerbaAlpina verwendeten und auf der Forschungsumgebung dargestellten sprachlichen Daten[2] bilden zunächst Sprachatlanten und dialektale Wörterbücher aus dem Untersuchungsgebiet, aus denen jeweils sprachliche Einzelbelege, phonetische und morphologische Typen[3] extrahiert und in die Datenbank (VA_DB) überführt werden. Voraussetzung dafür, dass Daten in VerbaAlpina übernommen werden, ist einzig deren Georeferenzierbarkeit: Sie müssen mindestens auf Ebene der einzelnen Gemeinde einem spezifischen Längen- und Breitengrad zugeteilt werden können (Krefeld – Lücke 2015b).
Die bereits vorliegenden Datenbestände werden von VerbaAlpina digitalisiert. Im Folgenden werden nun die einzelnen Schritte dieses Digitalisierungsprozesses beschrieben:

In einem ersten Schritt werden aus jeder Quelle (Sprachatlas, Wörterbuch usw.) diejenigen Stimuli bzw. Konzepte ausgewählt, die für die von VerbaAlpina bearbeiteten Sachgebiete relevant sind. Diese werden in entsprechenden Tabellen in der VA_DB gespeichert, in welcher auch die Verknüpfung der Stimuli mit den Konzepten abgebildet ist. Jeder Erhebungspunkt auf einer Karte in einem Sprachatlas wird in VA_DB durch Vergabe von Identifikatoren (»ID«s) eindeutig identifiziert und bleibt unverändert. Die auf diese Weise vorbereiteten Karten werden mit einem Webtool, dem von VerbaAlpina entwickelten Transkriptionstool (TT), transkribiert, mit dem Ziel, alle im jeweiligen Sprachatlas vorhandenen Informationen zu einem bestimmten Erhebungspunkt ohne Informationsverlust abzubilden. Dieser Vorgang ermöglicht den Zugriff auf die ursprüngliche Quellentranskription zu einem beliebigen Zeitpunkt. Das TT dient außerdem dazu, die Verknüpfung der Belege mit den entsprechenden Konzepten in der Datenbankstruktur abzubilden. Falls auf einer Karte mehrere Konzepte vorkommen, können mit Hilfe des TT die transkribierten Äußerungen mit einem anderen Konzept verknüpft werden als mit dem des jeweiligen Kartentitels (in VA_DB ‘Stimulus’ genannt).[4] Da je nach sprachwissenschaftlicher Tradition jeweils verschiedene Transkriptionssysteme verwendet werden bzw. wurden (z. B.: Böhmer-Ascoli, Rousselot-Gilliéron, Theuthonista), wird im Projekt der sogenannte *Betacode* (Krefeld – Lücke 2015a) eingesetzt, der auf den Zeichen

2 Darüber hinaus ist es möglich, verschiedene außersprachliche Daten, z. B. demographische oder historische, in die Datenbank einzuspeisen.
3 Mehr dazu in Krefeld – Lücke (2015d).
4 Bereits ein oberflächlicher Blick in den Sprach- und Sachatlas Italiens und der Südschweiz (AIS) zeigt, dass eine Karte (= VA Stimulus) oft mehrere Konzepte beinhalten kann.

aus dem ASCII-Bereich[5] basiert und mit jeder Tastatur ohne zusätzliche Software eingegeben werden kann. Die einheitliche und – im Vergleich zu den Ausgangstranskriptionen – einfachere Kodierung hat folgende Vorteile:

- schnellere Datenerfassung, die auch durch Personen die über keinerlei Vorkenntnisse von spezifischen Transkriptionssystemen verfügen, erfolgen kann;
- Erfassung beliebiger Zeichen und Diakritika, unabhängig davon, ob diese in Unicode[6] existieren oder nicht;
- kein Informationsverlust bezüglich Originaltranskription im jeweiligen Sprachatlas.

Unabhängig davon, wie der lautliche Wert einer transkribierten Einheit ist, werden alle gleich aussehenden Zeichen mit demselben Betacode-Zeichen kodiert (s. Abb. 1) und in der Datenbank gespeichert; Für jede Quelle existiert in der Datenbank eine sogenannte Codepage, mit Hilfe deren die Übersetzung in die IPA-Werte geleistet werden kann (vgl. Lücke 2015).

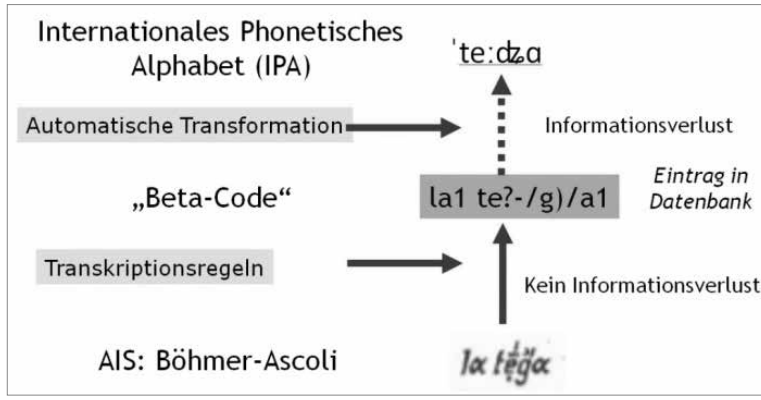

Abbildung 1: Betacode (Krefeld – Lücke 2015b)

Diese Verknüpfung verbindet den abgefragten Stimulus demnach mit einer konkreten Äußerung (= Einzelbeleg auf der Karte) und wird in der VA_DB festgehalten. Nach der vollständigen Transkription einer Karte werden die Äußerungen in einem nächsten Schritt tokenisiert, d. h. de facto in einzelne Wörter aufgeteilt und in einer separaten Datenbanktabelle abgespeichert[7]. Dieser Schritt ist notwendig,

5 S. v. ‚ASCII', in Fischer – Hofer (2008: 52).

6 S. v. ‚Unicode', in Fischer – Hofer (2008: 875).

7 Bei der Tokenisierung wird eine Äußerung in Einzeltokens zerlegt. So lautet beispielweise dieÄußerung zum Konzept SCHLAGSAHNE aus dem Atlante storico-linguistico-etnografico friulano (ASLEF) in Cergnéu (Gemeinde Nimis) *klabúk ot mlíéka* (klabúk ot mlíéka|Einzelbeleg|Cergnéu, com. di Nimis|ASLEF#3994|VA_15/1). Nach der Tokenisierung erhalten wir drei Tokens: 1. *klabuk*, 2. *ot*, 3. *mleka*, welche mit drei neuen Konzepten und drei morpho-lexikalischen Typen entsprechend verknüpft werden: 1. HUT – klobuk (sla. m.), 2. PRÄPOSITION – ot, 3. MILCH – mleko (sla. n.).

um morphosyntaktische Etikettierungen wortbezogen durchführen zu können, was wiederum für nachfolgende analytische Recherchen unabdingbar ist.

Auf die Tokenisierung folgt die Konvertierung der Daten in IPA zum Zweck einer großflächigen, quellenunabhängigen Datenvergleichbarkeit aller vorhandenen sprachlichen Belege. Für jede Datenquelle sind in einer Tabelle die Entsprechungen zwischen der Originaltranskription, dem Betacode und der IPA-Repräsentation dokumentiert.

Der nächste Schritt besteht in der Typisierung (vgl. Krefeld – Lücke 2015d) der entstandenen Tokens auf mehreren Ebenen: zunächst phonetisch und morpho-lexikalisch, dies sind zwei voneinander unabhängige Prozesse. Geplant ist, dass die phonetische Typisierung halbautomatisch erfolgt und sich nach zwei Algorithmen richten wird: dem Levenshtein-Algorithmus einerseits und dem SOUNDEX-Algorithmus andererseits[8]. Die morpho-lexikalische Typisierung erfolgt über ein von VerbaAlpina entwickeltes Typisierungstool (TypT) unter Berücksichtigung folgender Kategorien: *Sprache* (romanisch, germanisch, slawisch), *Wortart*, *Suffigierung*, *Genus*. Jeder morpho-lexikalische Typ wird durch eine standardsprachliche orthographische Form definiert, die, sofern möglich, wiederum passenden Lemmata der sogenannten *Referenzwörterbücher*[9] (zunächst auf Ebene der Standardsprache, dann auf jener des Dialekts) zugeordnet werden. Für den Fall, dass für einen spezifischen morpho-lexikalischen Typ keine entsprechende Schreibung belegt ist bzw. ein solcher Typ keinem Lemma eines der Referenzwörterbücher zugeordnet werden kann, wird mit den orthographischen Regeln der jeweiligen Standardsprache ein Lemma erstellt und dem zu typisierenden Einzelbeleg zugeordnet (diese Typen erhalten die Markierung ‚VA-Typ'). Darüber hinaus ermöglicht die kollaborative Konzeption des Projektes, dass verschiedene Benutzer des Portals einem sprachlichen Beleg verschiedene morpho-lexikalische Typen zuordnen. Alle Informationen werden in der VA_DB gespeichert und entsprechend markiert. Dem auf diese Weise erstellten morpho-lexikalischen Typ wird ein Basistyp, d. h. eine dem morpho-lexikalischen Typ zugrundeliegende lexikalische Basis, zugewiesen (vgl. Krefeld – Lücke 2015d).

Während der morpho-lexikalische Typ sprachgebunden ist, ist der Basistyp einzelsprachunabhängig. Die Verknüpfung der Ersteren mit dem sprachübergreifenden Basistyp erfolgt meist durch Etymologisierung. Nach der Bearbeitung aller Typisierungsebenen wird automatisch festgestellt, ob ein Sprachkontaktphänomen vorliegt, indem die jeweilige Sprachzuweisung (ger., rom., sla.) des morpho-lexikalischen Typs und des Basistyps verglichen werden: Fehlt eine

---

8 S. v. »soundex, SOUNDEX()« in Fischer – Hofer (2008: 784) und s. v. »Levenshtein-Distanz« in Fischer – Hofer (2008: 481).

9 Für die romanischen Belege werden Treccani und Le Trésor de la Langue Française Informatisé (TLFi) via Centre National de Ressources Textuelles et Lexicales (CNRTL) verwendet, für germanische Belege der Duden sowie das Idiotikon und für die slawischen Belege SSKJ. Vgl. dazu Krefeld – Lücke (2015c).

Übereinstimmung, kann dies ein Indiz für einen Sprachkontakt oder für ein gemeinsames Substrat sein.

Der Zugriff auf die so aufbereiteten Daten wird dem VerbaAlpina-Nutzer auf zweierlei Weise angeboten: Eine spezielle Schnittstelle (vap_), ermöglicht den direkten Zugang zu den relational strukturierten Daten über die VA_DB. Die interaktive Karte hingegen leistet eine frei kombinierbare Darstellung der Daten im Raum. Während die Textversion durch den Einsatz des Datenbankverwaltungssystems MySQL[10] sehr individuelle und komplexe Abfragen ermöglicht, ist die visuelle, kartographische Darstellung insofern vorteilhaft, als sie einen guten Überblick über die Gesamtheit der Daten vermittelt und zudem die Einbindung außersprachlicher Daten erlaubt. Daher sind die beiden Zugangswege als komplementär zu betrachten.

3 Datenlage, Datenvalidierung und Datenerhebung mittels Crowdsourcing

3.1 Datenlage

Aktuell befinden sich die Daten von 559 Karten (VA-Stimuli) aus verschiedenen Sprachatlanten des gesamten Alpenraums in der VA_DB [Stand: 25. 7. 2016]. Dabei wurden 1.091 Konzepte[11] extrahiert und 23.701 zugehörige Äußerungen nach der Transkription der entsprechenden Karten erfasst. Die Anzahl der Tokens in der Datenbank ist jedoch wesentlich höher, da pro Erhebungspunkt mehrere Belege registriert werden können (zusätzliche Belege, Mehrwortausdrücke, Satzkonstruktionen usw.). Die oben erwähnten Stimuli beziehen sich nur auf diejenigen Karten, die in der aktuellen Projektphase im Fokus von VerbaAlpina stehen, also solche aus dem Bereich Almwirtschaft. Es ist möglich und erwünscht, dass auch Stimuli außerhalb des aktuellen thematischen Fokus von VA transkribiert, typisiert und kartographisch dargestellt werden.[12] Bis dato [25. 7. 2016] brachte der Prozess der Typisierung 1.104 germanische, 610 ro-

---

10 S. v. ‚MySQL', in Fischer – Hofer (2008: 556).

11 Synonymisch zu Konzept kann hier ‚Sache' oder ‚Sachverhalt' verwendet werden. Nicht selten, kann eine Karte aus dem Sprachatlas mehrere Konzepte, meistens ein Überkonzept und mehrere Unterkonzepte, enthalten. Vgl. dazu AIS 1206 ‚zangola', die BUTTERFASS als Hauptkonzept und GERÄT ZUM BUTTERN, DURCH DREHEN; GERÄT ZUM BUTTERN, BARKENÄHNLICH, AN EINEM GESTELL AUFGEHÄNGT; GEFÄSS ZUM BUTTERN, RAHM WIRD IN OFFENEM GEFÄSS GESCHLAGEN; GEFÄSS ZUM BUTTERN, RAHM WIRD IN EINER FLASCHE ODER BLECHBÜCHSE GESCHÜTTELT als Unterkonzepte beinhaltet. Zum Prinzip der Konzeptbeschreibung vgl. Grimaldi – Krefeld 2015.

12 Wie bereits erwähnt ist die kollaborative Entwicklung des Projektes erwünscht. Daher steht die gesamte informatische Struktur jedem registrierten Nutzer zur Verfügung. Registrierte Nutzer können eigene Daten transkribieren, typisieren und kartographieren. VerbaAlpina ist für jegliche Vorschläge im Bereich der kollaborativen Entwicklung der Forschungsumgebung offen und nimmt diese gerne jederzeit entgegen.

manische und 210 slawische Typen hervor, die derzeit mit 251 lexikalischen Basistypen verknüpft sind. Beispielsweise emergieren für die lexikalische Basis ‚baita' folgende morpho-lexikalische Typen: *baita* (rom. f.), *baito* (rom. m.) *bait* (rom. m.), *baitun* (rom. m), *baitin* (rom. m.), *bajta* (sla. f.), *bajtica* (sla. f.) (vgl. Abbildung 2).

Die Anzahl der Konzepte, Äußerungen, Tokens und Typen wächst aktuell ständig, da die zuvor geschilderten Prozesse parallel ablaufen: Es wird kontinuierlich transkribiert, typisiert und etymologisiert.

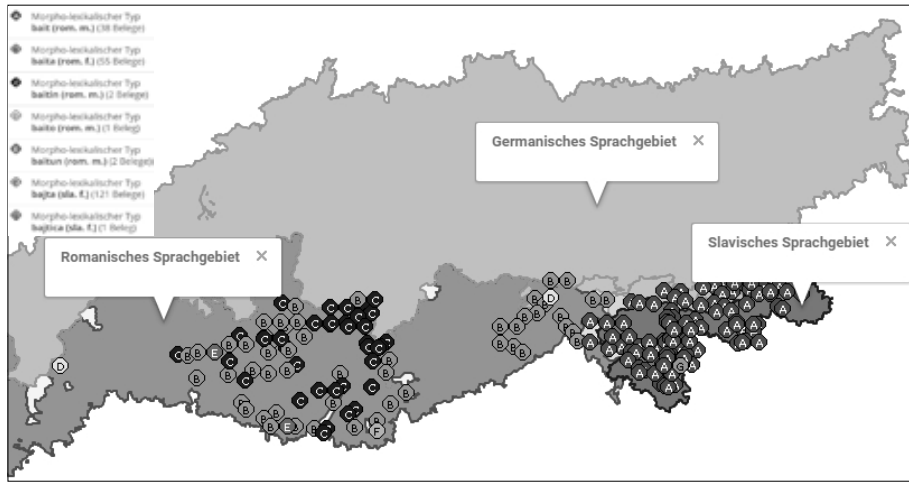

Abbildung 2: Geographische Verteilung der romanischen, slawischen und germanischen morpho-lexikalischen Typen mit der gleichen lexikalischen Basis ‚baita' (Die Abbildung wurde mit der interaktiven Karte generiert unter: https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=133 [letzter Zugriff: 14. 11. 2016].)

3.2 Die Frage der Datenqualität: das Problem der historischen Daten

Trotz der Gesamtheit des in Sprachatlanten und Wörterbüchern dokumentierten Sprachmaterials stellen sich methodische Fragen die sich auf die Qualität und Interpretierbarkeit der Daten beziehen. Beispielsweise zwingt die Konzeption der Sprachatlanten den Forscher dazu (vgl. Krefeld 2016), über die Generalisierbarkeit der Aussagen für einen Erhebungspunkt – und somit auch für den entsprechenden Dialekt – nachzudenken. In der Mehrzahl der Fälle findet man an einem einzelnen Aufnahmepunkt lediglich einen sprachlichen Beleg. Problematischer gestaltet sich die Situation, wenn an einem Erhebungspunkt zwei oder sogar drei unterschiedliche Äußerungen belegt sind. Dies lässt sich anhand von Belegen zum ASLEF-Stimulus ‚entrahmen' für den Ortspunkt Laglésie San Leopoldo anschaulich illustrieren: Die Belege *posnemat* (v. sla.)|morphTyp|Laglésie San Leopoldo|ASLEF#3995|VA_15/1 und *sbramare* (v. rom)|morphTyp|Laglésie San Leopoldo| ASLEF#3995|VA_15/1 gehören zwei verschiedenen Sprachfamilien

an. Der Erhebungspunkt wird durch den Informanten selbst als slawisch markiert. Folglich müssen folgende Fragen gestellt werden: Handelt es sich im Fall des rom. Einzelbelegs *sbramare* um einen Teil eines Soziolekts, eines Idiolekts oder stellt der Beleg ein Beispiel von *Codeswitching* eines zweisprachigen Sprechers dar?

Dass Dialekte unabhängig von der Dachsprache über die politischen Grenzen hinweg ihr Kontinuum bilden können, ist unumstritten (vgl. Chambers – Trudgill 1998: 5–7; Woolhiser 2005: 236–237). Sprachatlanten und dialektale Wörterbücher bleiben jedoch meist auf eine bestimmte Region oder gar einen einzelnen Ort begrenzt und können das dialektale Kontinuum bestenfalls teilweise berücksichtigen. Dadurch ist es traditionell kaum möglich, die Verbreitung von Phänomenen über die Dialekte bzw. sprachlichen Grenzen hinweg zu beobachten. Gerade durch die systematische grenzüberschreitende Dokumentation der Sprachdaten, wie sie von VerbaAlpina geleistet wird, ermöglicht nun genau das. Bei dieser systematischen Datenerfassung ergeben sich jedoch diverse Inkonsistenzen: Die verschiedenen Quellen (Sprachatlanten, Wörterbücher) dokumentieren für die unterschiedlichen Erhebungsgebiete nicht die jeweils identischen Konzepte, die Daten stammen aus unterschiedlichen Zeiten, die Erhebungsnetze sind unterschiedlich dicht. Solche Inkonsistenzen zu beheben, Datenlücken zu schließen und bestehende Daten zu validieren ist unter anderem Ziel von VerbaAlpina. Dieses soll mithilfe des sogenannten Crowdsourcings erreicht werden.

3.3 Die Konzeption der Neuerhebungen mithilfe von Crowdsourcing in VerbaAlpina

In einer im Rahmen des Projektes entstandenen Bachelorarbeit (Grimaldi 2016a) wurde eine erste Version eines Tools für die Datenerhebung und -validierung unter Einbezug der *Crowd* entwickelt (VACCA)[13]. Diese Anwendung, die sowohl vom Handy als auch vom Computer aus über eine Webseite aufgerufen werden kann, befindet sich momentan in der Pretestphase. Die Grundfunktionalität besteht darin, jene Konzepte aus der VA-Datenbank abzufragen, zu welchen die oben genannten Sprachatlanten keine oder inkonsistente Daten liefern. Eine weitere Funktion besteht darin, Konzepte abzufragen, für die bereits vorhandenen Belege validiert werden müssen. Das Erfragen solcher Belege zu bestimmten Konzepten wird, wenn möglich, durch Fotografien oder Videoaufnahmen aus der projekteigenen Mediathek[14] unterstützt. Der Benutzer wird zunächst aufgefordert, den eigenen Dialekt bzw. Wohnort zu benennen und 20 Wörter zu verschiedenen Konzepten, in zwei Serien mit jeweils zehn Wörtern, anzugeben. Die Wahl der Verschriftung der Wörter wird dem Nutzer überlassen.

---

13 VerbaAlpina Cooperative Crowdsourcing Application.

14 Die Mediathek ist eine Sammlung von Fotografien, Videoaufnahmen und anderen digitalen Dokumenten, die für VerbaAlpina relevant sind und auf dem Web-Server des Projektes aufbewahrt werden. Die einzelnen digitalen Dateien sind georeferenziert und mit den entsprechenden Konzepten aus der Datenbank verknüpft.

Zur Bestimmung der Gewichtung der eingegebenen Daten ist ein sogenannter Kompetenztest Teil der Befragung, mit dessen Hilfe durch gezielte Wissensabfragen die sprachliche und/oder sachliche Kompetenz der Personen eingeschätzt wird. (vgl. Krefeld 2015a). Nach den 20 Fragen wird der Informant nach seinen sozio-demographischen Daten gefragt (Geschlecht, Alter) und ob er sich auf der Projektwebseite registrieren möchte. Darüber hinaus wird die Frage gestellt, ob er zusätzlich zur bereits erfolgten Eingabe im eigenen Dialekt weitere Daten zu den gerade abgefragten Konzepten aus einem anderen Dialekt bzw. von einer anderen Ortschaft eintragen möchte.

3.4 Crowdsourcing als Möglichkeit von linguistischen Erhebungen

Crowdsourcing ist im Bereich der Sprachwissenschaft nicht neu. Bereits am Anfang des 20. Jahrhunderts startete James Murray eine Kampagne zur Erhebung von Daten für den Oxford English Dictionary (OED). Er richtete dafür an die Menschen aus dem ganzen britischen Königreich die Bitte, auf kleinen Zetteln Wörter und Beispiele für die Anwendung derselben aufzuschreiben und diese per Post an ihn zu senden. Daraus resultierte die erste crowdbasierte Ausgabe des OED (vgl. Salazar 2014). Die Rolle des Sprechers beschränkte sich also bereits zu diesem frühen Zeitpunkt nicht mehr auf jene eines reinen Untersuchungsobjekts, sondern gestaltete sich aktiver als bis anhin, beispielsweise bei Erhebungen für Sprachatlanten – das Crowdsourcing *avant la lettre* war geboren. Ungefähr 100 Jahre später entfaltet das Verfahren mit der Erfindung des Internets nun sein eigentliches Potential im linguistischen Kontext.

Im Allgemeinen wird Crowdsourcing als eine Methode der webbasierten Datenerhebung beschrieben (vgl. Juska-Bacher – Biemann – Quasthoff 2013: 14–20) und gewinnt für die linguistische Forschung immer mehr an Bedeutung (vgl. Munro u. a. 2010: 122). Verschiedene Projekte werden dafür eingesetzt, bedrohte Sprachen oder Varietäten zu dokumentieren (z. B. beim Rosetta Project[15]), Texte zu transkribieren (z. B. Shakespeare‘s World[16]), anaphorische Beziehungen im Text zu identifizieren (z. B. Phrase Detectives[17]) oder auch neue Daten zu erheben (DialektÄpp[18], VerbaAlpina). Wie die *Crowd* in die Forschung und somit auch in den Prozess der Datenerhebung und -bearbeitung einbezogen wird, hängt von der Art des Projektes ab. Abgesehen von den Vorteilen, wie der Kosten- und Zeitersparnis, der Möglichkeit der großflächigen Datenerhebung und teilweise auch der Bearbeitung der Daten durch die *Crowd* sowie dem damit verbundenen Paradigmenwechsel (vgl. Krefeld 2015e), gilt es beim Einsatz von Crowdsourcing, auch einige Schwierigkeiten zu bedenken:

---

15 Vgl. http://rosettaproject.org/blog/02013/apr/17/android-app-language-documentation/ [letzter Zugriff: 2. 8. 2016].

16 Vgl. https://www.shakespearesworld.org/#/ [letzter Zugriff: 2. 8. 2016].

17 Vgl. http://anawiki.essex.ac.uk/phrasedetectives/ [letzter Zugriff: 2. 8. 2016].

18 Vgl. http://dialaektaepp.ch/ [letzter Zugriff: 2. 8. 2016].

Eine VerbaAlpina typische Herausforderung von Crowdsourcing besteht in der Technologie selbst. In Abhängigkeit davon, welche Sprechergemeinschaft bzw. welcher Sprachbereich erforscht werden soll, kann das gewählte technologische Verfahren ein Hindernis darstellen. Grimaldi (2016b) bemerkt in Bezug auf VerbaAlpina Folgendes:

> Zunächst stellt die Erreichbarkeit der Personen, die interessant für VerbaAlpina sind, ein gewisses Problem dar. Ein echter Experte bzgl. Milch- und Käseproduktion und daher für den Alpenwortschatz im Allgemeinen z. B. ein Senn, ist tendenziell ein älterer Mensch, der im VerbaAlpinagebiet lebt. Das heißt, es könnte schwierig werden, jene Experten zum einen altersbedingt mit Social Media, App oder Website zu erreichen. Zum anderen könnte die zweite Schwierigkeit geographischer Natur sein, da der Internetanschluss im Gebirge und somit die Zugänglichkeit nicht abzusehen ist. Auch bei schlechtem Empfang o. Ä. könnten Komplikationen für das Nutzen der App/Website auftreten. [Hervorhebungen A. W.]

Das Problem bezüglich des Alters der Informanten stellte sich auch Leemann u. a. (2016: Abb. 14) bei ihren appbasierten Crowdsourcing-Erhebungen: Bei 16.000 Teilnehmern ist die Gruppe der 11- bis 30-Jährigen am stärksten vertreten, während die Partizipation ab dem Alter von 40 Jahren proportional abnimmt.

Für VerbaAlpina gilt es zudem zu bedenken, dass im Erhebungsgebiet die Erreichbarkeit via Internet unter Umständen nicht jederzeit gegeben ist, so z. B. wenn die Erhebung per Handy bei schlechter Netzabdeckung oder während der Almsaison, in der die möglichen Gewährspersonen zeitlich stark beansprucht sind, stattfinden soll.

3.5 Crowdsourcing für die Benutzer attraktiv gestalten

Die für einen nicht-linguistischen Kontext geschaffene Definition von Crowdsourcing von Estellés-Arolas – González-Ladrón-de-Guevara (2012: 9 f.) gibt Hinweise darauf, welche Aspekte bei der Planung des Einsatzes von Crowdsourcing grundsätzlich berücksichtigt werden müssen, um eine möglichst große Anzahl von Informanten zu gewinnen:

> Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task; of variable complexity and modularity, and; in which the crowd should participate, bringing their work, money, knowledge **[and/or]** experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that which the user has brought to the venture, whose form will depend on the type of activity undertaken. [Hervorhebungen A. W.]

Ausgehend davon lässt sich also für den linguistischen Kontext konstatieren, dass Crowdsourcing auf der freiwilligen Mitarbeit des Teilnehmers (Laie/Wissenschaftler), der in verschiedenen Rollen auftreten kann, basiert, dass dieser über sprachliche Informationen und/oder Fähigkeiten verfügt und bereit ist, diese zu teilen. Die Arbeit des *Crowders*, also als Teil der *Crowd*, ist für ihn auf eine vom

Design des Projektes definierte Art und Weise *immer* profitabel (nicht unbedingt in Form eines monetären Profits) und für das Projekt nützlich.

Da die *Crowd* eine sehr heterogene Gruppe darstellt (vgl. Reips u. a. 2015: 4), sind ihre Kompetenzen, ihr Wissen und ihre Fähigkeiten hinsichtlich eines bestimmten Lebensbereiches (im Fall von VerbaAlpina im Bereich der Milchwirtschaft) ungleich auf die Einzelpersonen verteilt.

Um das große Informationspotential der Forschungsteilnehmer zu entfalten, wäre es jedoch sinnvoll, Einzelpersonen nach projektspezifischen Kategorien zu gruppieren und ihnen zudem die Möglichkeit zu geben, ihre Teilnahme in einem vorgesehenen Rahmen selbst zu gestalten.

Diese Betrachtung der Rolle des einzelnen Teilnehmers, die sich im Projektverlauf weiterentwickeln könnte, und die man als evolutiv bezeichnen kann, wurde in der ersten Version von VACCA nicht in Erwägung gezogen. So könnte man beispielsweise jeder Gewährsperson ermöglichen, einen „Karrierepfad" zu gehen, also den eigenen Weg im und für das Projekt selbst mitzugestalten. Jeder Teilnehmer beginnt als reiner Informant auf der Webseite, jedoch mit der Option, seinen Status nach der Registrierung zu verbessern. Nach der Anmeldung steht es ihm frei, selbst zu entscheiden, welche Aufgaben er bearbeiten möchte (Belohnung: Der Informant fühlt sich als „etwas Besonderes" (*self-esteem*) und gehört zu einer anderen Gruppe als die Masse (*social recognition*). Er ist Teil der Gemeinschaft von Menschen, die sich mit dem Thema „Alpendialekte, Milchverarbeitung" befassen und besonders gut auskennen.

Die Differenzierung – die projektintern geschehen kann – soll jedoch an dieser Stelle nicht enden: Es muss nämlich unter den registrierten Nutzern weiter nach den sogenannten guten *Crowdern* gesucht werden. Deren Rolle lässt sich mit dem soziologischen Begriff *Gatekeeper* (vgl. dazu Atkinson – Hammersley 2003: 49–53) umschreiben, da sie im Sinne der perzeptiven Varietätenlinguistik das Problem der Erreichbarkeit von älteren Menschen und somit auch von sprachlichen und sachlichen Informationen lösen können. Laut Krefeld – Pustka (2010b: 10–13) verfügt jeder Sprachbenutzer über Informationen zu/r eigener/n Sprache/n und zu der/n Sprache/n der anderen. Dieses Sprachmaterial basiert auf der Perzeption der Sprecher und kann, wie in verschiedenen Studien in Krefeld – Pustka (2010a) bestätigt, von Menschen jederzeit abgerufen werden. Wir sprechen hier von mehreren Dimensionen des Wissens bei einem einzelnen Sprecher, nämlich von der Synchronie und der Mikrodiachronie. Ein einziger Sprachbenutzer kann demnach folgendes Sprachmaterial liefern:

(a) Dialektwörter aus dem eigenem Repertoire;
(b) Dialektwörter aus dem Repertoire der anderen Sprecher;
(c) Dialektwörter aus dem Repertoire der eigenen Sprachumgebung, der jüngeren oder der älteren Menschen, relativ zum eigenen Alter;
(d) Dialektwörter aus dem Repertoire der externen Sprachumgebung, der jüngeren oder der älteren Menschen, relativ zum eigenen Alter.

Die Kriterien wie die Anzahl der bearbeiteten Konzepte, die Qualität der Daten (diese ergibt sich aus dem Vergleich mit der allgemeinen Abstimmung und Überprüfung durch Experten), die Aktivität im Projekt und auch der Ausbau des eigenen Profils (durch Vervollständigung der persönlichen Informationen usw.) sollen auf der einen Seite den *Crowder* (sei er Laie oder Wissenschaftler) motivieren, sich gründlicher mit dem Projekt zu befassen. Auf der anderen Seite kann das Projekt die Charakteristika des Einzel*crowders* beobachten und ihm andere Aufgaben erteilen als die für den Rest der Teilnehmenden vorgesehenen. Es wäre sogar denkbar, Einzelpersonen für bestimmte Aufgaben (die etwas komplizierter sind) zu schulen, beispielsweise für die Ausbildung von Exploratoren, die neue Daten für das Projekt erheben. Dies soll umgekehrt aber nicht heißen, dass die nicht angemeldeten Nutzer ausgeschlossen werden und ihre Eingaben nicht in den Datenbestand einfließen. Im Gegenteil: Daten von verschiedenen Typen von Informanten werden in der Datenbank abgespeichert aber unterschiedlich gewichtet. Die vom Projekt als ‚unsichere' Daten (die z.B. von nicht-registrierten Nutzern stammen) markierten Einträge können und müssen der *Crowd* und ggf. den Experten zur Validierung gestellt werden.

4 Abschliessende Anmerkungen

Unter Berücksichtigung der Ziele von VerbaAlpina, der oben beschriebenen methodischen Probleme von traditionellen Datenquellen und der Überlegungen bezüglich Crowdsourcing wird im Projekt gegenwärtig ein auf der ersten Version von VACCA basierendes Backend für registrierte Informanten entwickelt. Vorgesehen ist, dort für jeden angemeldeten Nutzer folgende Optionen zur Verfügung zu stellen:

- Auswahl der inhaltlichen Bereiche (Konzeptabfragen zu selbstausgewählten Themenbereichen, Datenvalidierung), die bearbeitet werden können;
- Übersicht über die bereits gelieferten Daten (eigene und fremde);
- Übersicht über den Fortschritt des Projektes;
- Hochladen und Georeferenzieren von eigenen Mediendateien und deren Verknüpfung mit Konzepten;
- Zugang zu einer Austauschplattform zwischen Wissenschaftlern, Laien und Projekt;
- Informationen über mögliche andere Typen der Kooperation im Rahmen des Projektes;
- Einbindung in die sozialen Netzwerke.

Die aktive und dynamische Einbindung verschiedener Nutzer kann als Faktor für eine gesteigerte intrinsische Motivation und das Aufkommen des Gefühls der Projektzugehörigkeit fungieren, die sich positiv auf eine langfristige Kooperation

mit dem Projekt auswirken können. Sowohl für die Motivation des Nutzers als auch für die Verbreitung der Informationen und Forschungsergebnisse des Projektes wäre folglich die Einbindung der sozialen Netzwerke zu überlegen. Um die Qualität und Konsistenz der Crowdsourcing-Daten zu gewährleisten, muss eine Datenvalidierung auf mehreren Ebenen implementiert und in regelmäßigen Abständen optimiert werden.

Trotz aller Komplexität im Umgang mit Crowdsourcing sind dessen Vorteile – insbesondere für eine großräumige Forschung – nicht zu übersehen. Mit diesem Tool beabsichtigen wir vor allem die Lücken der historischen Daten im Alpenraum zu schließen und neue Daten dazu zu erheben, die systematisch aufbereitet und strukturiert in der Datenbank abgebildet werden.

BIBLIOGRAPHIE

AIS = Karl Jaberg – Jakob Jud, *Sprach- und Sachatlas Italiens und der Südschweiz* 1–7, Zofingen: Ringier, 1928–1940.

ASLEF = Giovan Battista Pellegrini, *Atlante storico-linguistico-etnografico friulano* 1–6, Padova: Istituto di Glottologia e Fonetica dell'Università die Padova u. a., 1974–1986.

Atkinson – Hammersley 2003 = Paul Atkinson – Martyn Hammersley, *Ethnography: principles in practice*, London – New York: Routledge, ^{3}2003.

Chambers – Trudgill 1998 = Jack K. Chambers – Peter Trudgill, *Dialectology*, Cambridge: Cambridge University Press, ^{2}1998.

CNRTL = *Centre National de Ressources Textuelles et Lexicales*, http://www.cnrtl.fr/ [letzter Zugriff: 17. 8. 2016].

Dialekt Äpp = http://dialaektaepp.ch/ [letzter Zugriff: 2. 8. 2016].

Estellés-Arolas – González-Ladrón-de-Guevara 2012 = Enrique Estellés-Arolas – Fernando González-Ladrón-de-Guevara, Towards an Integrated Crowdsourcing Definition, *Journal of Information Science* 38 (2012), Nr. 2, 189–200, doi:10.1177/0165551512437638.

Fischer – Hofer 2008 = Peter Fischer – Peter Hofer, *Lexikon der Informatik*, Berlin – Heidelberg: Springer, ^{14}2008.

Grimaldi 2016a = Giorgia Grimaldi, *»Come dici...?«: Crowdsourcing-Pretest für VerbaAlpina mittels einer App/mobilen Website*, Bachelorarbeit, Ludwig-Maximilians-Universität München, 2016, https://www.verba-alpina.gwi.uni-muenchen.de/?p=2465 [letzter Zugriff: 18. 8. 2016].

Grimaldi 2016b = Giorgia Grimaldi, Methodik und Vorgehensweise: Konzeption der Personenrecherche für die Crowd, Kapitel 3, *»Come dici...?«: Crowdsourcing-Pretest für VerbaAlpina mittels einer App/mobilen Website*, von Giorgia Grimaldi, Bachelorarbeit, Ludwig-Maximilians-Universität München, 2016, https://www.verba-alpina.gwi.uni-muenchen.de/?p=2825 [letzter Zugriff: 18. 8. 2016].

Grimaldi – Krefeld 2015 = Giorgia Grimaldi – Thomas Krefeld, s. v. ‚Konzeptbeschreibung', VA-de 16/1, Methodologie, https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=161&letter=K#38.

Hubschmid 1951 = Johannes Hubschmid, *Alpenwörter romanischen und vorromanischen Ursprungs*, Bern: Francke, 1951.

Juska-Bacher – Biemann – Quasthoff 2013 = Britta Juska-Bacher – Chris Biemann – Uwe Quasthoff, Webbasierte linguistische Forschung: Möglichkeiten und Begrenzungen beim Umgang mit Massendaten, *Linguistik online* 61 (2013), Nr. 4, 7–30.

Krefeld 2015a = Thomas Krefeld, s. v. ‚Crowdsourcing', VA-de 16/1, Methodologie, https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=161&letter=C#12.

Krefeld 2015b = Thomas Krefeld, s. v. ‚interlinguale Geolinuistik', VA-de 16/1, Methodologie, Methodologie, https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=161&letter=I#32.

Krefeld 2015c = Thomas Krefeld, s. v. ‚Publikation', VA-de 16/1, Methodologie, https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=161&letter=P#47.

Krefeld 2015d = Thomas Krefeld, s. v. ‚Stratigraphie', VA-de 16/1, Methodologie, https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=161&letter=S#55.

Krefeld 2015e = Thomas Krefeld, s. v. ‚Wissenschaftskommunikation im Web', VA-de 16/1, Methodologie, https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=161&letter=W#62.

Krefeld 2016 = Thomas Krefeld, *Geolinguistik in der Perspektive der ‚digital humanities' (am Beispiel von Verba Alpina)*, https://www.dh-lehre.gwi.uni-muenchen.de/?lehrveranstaltung=geolinguistik-in-der-perspektive-der-digital-humanities-am-beispiel-von-verba-alpina-2 [letzter Zugriff: 25. 7. 2016].

Krefeld – Lücke 2014 = Thomas Krefeld – Stephan Lücke, *VerbaAlpina* – Der alpine Kulturraum im Spiegel seiner Mehrsprachigkeit, *Ladinia* 38 (2014), 189–219.

Krefeld – Lücke 2015a = Thomas Krefeld – Stephan Lücke, s. v. ‚Betacode', VA-de 16/1, Methodologie, https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=161&letter=B#7.

Krefeld – Lücke 2015b = Thomas Krefeld – Stephan Lücke, s. v. ‚Georeferenzierung', VA-de 16/1, Methodologie, https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=161&letter=G#27.

Krefeld – Lücke 2015c = Thomas Krefeld –Stephan Lücke, s. v. ‚Referenzwörterbücher', VA-de 16/1, Methodologie, https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=161&letter=R#5.

Krefeld – Lücke 2015d = Thomas Krefeld – Lücke Stephan, s. v. ‚Typisierung', VA-de 16/1, Methodologie, https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=161&letter=T#58.

Krefeld – Pustka 2010a = Thomas Krefeld – Elissa Pustka (Hrsg.), *Perzeptive Varietätenlinguistik*, Frankfurt am Main u. a.: Peter Lang, 2010 (Spazi comunicativi = Kommunikative Räume 8).

Krefeld – Pustka 2010b = Thomas Krefeld – Elissa Pustka, Für eine perzeptive Varietätenlinguistik, *Perzeptive Varietätenlinguistik*, hrsg. von Thomas Krefeld – Elissa Pustka, Frankfurt am Main u. a.: Peter Lang, 2010 (Spazi comunicativi = Kommunikative Räume 8), 9–28.

Leemann u. a. 2016 = Adrian Leemann u. a., Crowdsourcing Language Change with Smartphone Applications, *PloS one* 11 (2016), Nr. 1, DOI:http://dx.doi.org/10.1371/journal.pone.0143060.

Lücke 2015 = Stephan Lücke, s. v. ‚Codepage', VA-de 16/1, Methodologie, https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=161&letter=C#11.

Minnich 1989 = Robert Gary Minnich, At the interface of the Germanic, Romance and Slavic worlds: folk culture as an idiom of collective self-images in Southeastern Alps, *Studia Ethnologica* 2 (1989), 163–180.

Munro u. a. 2010 = Robert Munro u. a., Crowdsourcing and language studies: the new generation of linguistic data, *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, Los Angeles, 6. 6. 2010, The Association for Computational Linguistics: Stroudsburg, PA, 2010, 122–130, http://www.aclweb.org/anthology/W/W10/W10-0719.pdf [letzter Zugriff 17. 8. 2016].

Phrase Detectives = http://anawiki.essex.ac.uk/phrasedetectives/ [letzter Zugriff: 2. 8. 2016].

Reips u. a. 2015 = Ulf-Dietrich Reips u. a., Methodological challenges in the use of the Internet for scientific research: Ten solutions and recommendations, *Studia Psychologica: Advance online publication*, 2015.

Rosetta Project = http://rosettaproject.org/blog/02013/apr/17/android-app-language-documentation/ [letzter Zugriff: 2. 8. 2016].

Salazar 2014 = Danica Salazar, How can World Englishes benefit from crowdsourcing?, *Oxford-Words blog*, 2014, http://blog.oxforddictionaries.com/2014/02/can-world-englishes-benefit-crowdsourcing/ [letzter Zugriff: 17. 8. 2016].

Shakespeare's World = https://www.shakespearesworld.org/#/ [letzter Zugriff: 2. 8. 2016].

SSKJ = *Slovar slovenskega knjižnega jezika*, Ljubljana: ZRC SAZU, 2000, http://bos.zrc-sazu.si/sskj.html [letzter Zugriff: 17. 8. 2016].

TLFi = *Trésor de la Langue Française informatisé*, http://www.atilf.fr/ [letzter Zugriff: 17. 8. 2016].

Treccani = *Vocabolario Treccani*, http://www.treccani.it/vocabolario/ [letzter Zugriff: 17. 8. 2016].

VerbaAlpina = VerbaAlpina (VA), http://www.verba-alpina.gwi.uni-muenchen.de, 15/1.
Woolhiser 2005 = Curt Woolhiser, Political borders and dialect divergence/convergence in Europe, *Dialect Change: Convergence and Divergence in European Languages*, hrsg. von Peter Auer – Frans Hinskens – Paul Kerswill, Cambridge u. a.: Cambridge University Press, 2005, 236–262.

POVZETEK

Pomen in vloga množičnega zunanjega izvajanja v projektu VerbaAlpina

Prispevek predstavlja spletni projekt VerbaAlpina, ki poteka na Univerzi Ludvika in Maksimilijana v Münchnu in ki ima za cilj sistematično analizo alpskega besedja v zvezi z realnostjo, ki je tipično za Alpe in ki prehaja jezikovne meje. Projekt poleg želenega dokumentiranja jezikov iz alpskega prostora (dokumentiranje ima za osnovo zgodovinske vire, kot so jezikovni atlasi, narečni slovarji in številni digitalni projekti) predstavlja virtualno raziskovalno okolje, ki je mišljeno tako za znanstvenike kot tudi za laike, ki se zanimajo za alpske jezike. V prispevku so opisani posamezni koraki, ki so potrebni pri obdelavi zgodovinskih podatkov, da bi se jih dalo strukturirano predstaviti v podatkovni zbirki. Tako na primer sistem Betacode primeljivo predstavlja različne transkripcijske sisteme, pač glede na znanstveno tradicijo. Orisani so izzivi, ki se tičejo prenosa in primerjave podatkov. Ker zgodovinski podatki izkazujejo vrzeli in nedoslednosti, si sodelujoči pri projektu prizadevajo kar se da obširno zajeti nove podatke in jih povezati s starimi. V ta namen projekt uporablja množico sodelujočih in orodja za množično zunanje izvajanje (crowdsourcing).