Erschienen in: Doval, Irene/Liste Lamas, Else (Hrsg.): Germanistik im Umbruch - Linguistik, Übersetzung und DaF. - Berlin: Frank & Timme, 2019. S. 13-25. (Germanistik International 7)

# Herausforderungen eines innovativen lexikografischen Projekts zu Besonderheiten des gesprochenen Deutsch in der Interaktion

MEIKE MELISS / CHRISTINE MÖHRS

*Universidade de Santiago de Compostela / Leibniz-Institut für Deutsche Sprache*

**Abstract**

The article presents the scientific and methodological challenges for the development of an innovative, corpus-based lexicographic resource for the lexicon of spoken German in interaction and points out new ways for lexicographical work. In addition to general project information on the starting points, the data basis, the methods, objectives and the concrete subject area, selected results of two project-related empirical studies on expectations on a lexicographic resource of spoken German are outlined. For corpus-based quantitative information, the possibilities of a tool developed in the framework of the project are presented. Furthermore, an insight into the conceptual and methodological considerations with regard to the microstructure of the lexicographical resource is provided.

**Keywords:** lexicology, interaction, spoken German, lexicography, corpus-based

**Zusammenfassung**

Der Beitrag stellt die wissenschaftlichen und methodologischen Herausforderungen für die Erstellung einer innovativen, korpusbasierten lexikografischen Ressource zur Lexik des gesprochenen Deutsch in der Interaktion vor und zeigt neue Wege für lexikografische Arbeiten auf. Neben allgemeinen Projektinformationen zu den Ausgangspunkten, der Datengrundlage, den Methoden, Zielen und dem konkreten Gegenstandsbereich werden ausgewählte Ergebnisse von zwei projektbezogenen empirischen Studien zu Erwartungshaltungen an eine lexikografische Ressource des gesprochenen Deutsch präsentiert. Für korpusbasierte quantitative Informationen werden die Möglichkeiten eines Tools, welches im Rahmen des Projekts entwickelt wurde, aufgezeigt. Außerdem wird ein Einblick in die konzeptionellen und methodologischen Überlegungen zur Mikrostruktur der geplanten Ressource gegeben.

**Schlüsselwörter:** Lexikologie, Interaktion, gesprochenes Deutsch, Lexikografie, korpusbasiert

## 1. Einleitung

Ziel des Beitrages[1] ist es, die wissenschaftlichen und methodologischen Herausforderungen für die Erstellung einer korpusbasierten lexikografischen Ressource zur Lexik des gesprochenen Deutsch im Rahmen des Projekts LeGeDe[2] vorzustellen. Zunächst werden allgemeine Projektinformationen zu den Ausgangspunkten, zum Gegenstandsbereich, zu den Methoden und Zielen sowie zu der Datengrundlage skizziert. Die Ausgangspunkte sollen außerdem in Zusammenhang mit ausgewählten Ergebnissen aus zwei projektbezogenen empirischen Studien zu Erwartungshaltungen an eine lexikografische Ressource des gesprochenen Deutsch betrachtet werden (vgl. Abschnitte 2 und 3).[3] Schließlich wird von einigen konzeptionellen und methodologischen Überlegungen sowie von zukunftsgerichteten Betrachtungen zur lexikografischen

[1] Wir danken an dieser Stelle Annette Klosa-Kückelhaus für die Durchsicht des Manuskriptes und ihre vielen wertvollen Hinweise.

[2] Das Forschungsprojekt „Lexik des gesprochenen Deutsch" (LeGeDe) ist ein drittmittelgefördertes Projekt der Leibniz-Gemeinschaft, wird als Kooperationsprojekt der Abteilungen Pragmatik und Lexik am Leibniz-Institut für Deutsche Sprache (IDS) in Mannheim umgesetzt und hat eine Laufzeit von drei Jahren (2016-2019). Eine detaillierte Projektbeschreibung findet sich in Meliss/Möhrs (2017).

[3] Informationen zu den Rahmenbedingungen beider Umfragen sind auf den LeGeDe-Projektseiten zusammengestellt, außerdem können dort auch die Fragebögen abgerufen werden: <http://www.ids-mannheim.de/lexik/lexik-

Umsetzung berichtet (vgl. Abschnitte 4 und 5). Auch diese Überlegungen werden in Teilen mit Ergebnissen aus den oben genannten empirischen Studien in Verbindung gesetzt.

## 2. Ausgangspunkte und Hauptannahmen

Das LeGeDe-Projekt verfolgt das Ziel, eine korpusbasierte elektronische Ressource, die lexikalische Besonderheiten des gesprochenen Deutsch in der Interaktion zum Gegenstand hat, zu entwickeln.[4] Im Rahmen der Projektarbeit sollen mit unterschiedlichen korpusbasierten Methoden v. a. diejenigen Spezifika der gesprochensprachlichen Lexik in der Interaktion identifiziert, analysiert und beschrieben werden, die in der bisherigen lexikografischen Kodifizierung und Forschung eher vernachlässigt wurden und werden. Die Entwicklung neuartiger lexikografischer Beschreibungsformate und Angabetypen, die Formen, Bedeutungen und Funktionen von lexikalischen Einheiten in interaktionalen Kontexten besonders in den Blick nehmen, ist dabei ein wichtiger Teil der Konzeption des geplanten Wörterbuches (= WB[5]). Der anvisierte lexikografische Prototyp soll als mögliche Zielgruppe v. a. Gesprächsforscher, Lexikologen, Sprachwissenschaftler und Sprachlehrende ansprechen (vgl. dazu Ergebnisse aus den Umfragen Meliss/Möhrs/Ribeiro Silveira 2018: 124f. und 132; Meliss/Möhrs/Ribeiro Silveira 2019: 116).

Das LeGeDe-Projekt basiert auf folgenden vier Hauptannahmen und Beobachtungen: (i) Es existieren Unterschiede in der Lexik des gesprochenen im Vergleich zum geschriebenen Deutsch auf unterschiedlichen Ebenen (vgl. Deppermann et al. Hg. 2017; Fiehler 2016; Imo 2007; Schwitalla [4]2012). (ii) Die lexikografische Kodifizierung der interaktionstypischen Besonderheiten der gesprochensprachlichen Lexik des Deutschen ist unzureichend (vgl. u. a. Meliss 2016; Meliss/Möhrs 2018; Meliss/Möhrs/Ribeiro Silveira 2019; Moon 1998; Trap-Jensen 2004). Der Vergleich des Informationsangebotes aus verschiedenen lexikografischen Standardwerken[6] lässt die oben genannten Schlüsse zu, sodass es gilt, die bestehenden Desiderata für diesen Gegenstandsbereich aufzudecken und neue korpusbasierte Methoden für die Erstellung einer lexikografischen Ressource zu erarbeiten, die die vorhandenen Lücken schließen möchten. (iii) Es ist außerdem zu beobachten, dass in den letzten Jahren der Informationsbedarf zu typisch gesprochensprachlicher Lexik allgemein und in unterschiedlichen Anwendungsbereichen, wie z. B. in Unterricht und Lehre (speziell im Sekundarbereich und in den Bereichen Deutsch als Fremd- und/oder Zweitsprache) sowie im Verlagswesen in Verbindung mit der Erstellung von geeigneten Unterrichtsmaterialien gestiegen ist (vgl. Handwerker et al. Hg. 2016; Imo/Moraldo Hg. 2015; Moraldo/Missaglia Hg. 2013; Reeg et al. Hg. 2012; Sieberg 2013). Auch im „Gemeinsamen europäischen Referenzrahmen für Sprachen" (= GeR) wird u. a. zum Beurteilungsraster zur mündlichen Kommunikation und dem Parameter „Interaktion" für Niveau C1 explizit darauf hingewiesen, dass der Lernende „[…] aus einem ohne weiteres verfügbaren Repertoire von Diskursmitteln eine geeignete Wendung auswählen [kann], um

---

des-gesprochenen-deutsch/projektbeschreibung/empirische-forschung.html>. Ergebnisse zu den Studien allgemein finden sich in Meliss/Möhrs/Ribeiro Silveira (2018), mit L2-Perspektive in Meliss/Möhrs/Ribeiro Silveira (2019) sowie in Meliss/Möhrs (2018).

[4] Am IDS sind in den letzten Jahren die Kompetenzen gereift, die es ermöglichen, ein solches empirisch fundiertes Wörterbuch des gesprochenen Deutsch zu erstellen. Es liegen hinreichend große Korpora des gesprochenen Deutsch, Expertise in der Konzeption und Realisierung komplexer multimedialer Internetwörterbücher und Erfahrungen in der lexikologischen und semantischen Analyse gesprochener Sprache in der Interaktion vor.

[5] Im Folgenden wird der Ausdruck „Wörterbuch" (in allen Flexionsformen) mit „WB" abgekürzt.

[6] Wir verweisen hier exemplarisch auf die lexikografischen Einträge zu *gucken* und *okay* in LGWB-DaF, DWDS und Duden-online.

seine/ihre Äußerung angemessen einzuleiten, wenn er/sie das Wort ergreifen oder behalten will, oder um die eigenen Beiträge geschickt mit denen anderer Personen zu verbinden" (Trim et al. 2001: 37). Die Lernenden können den formulierten Ansprüchen des GeR und der Lehrpläne nur schwerlich gerecht werden, wenn keine Ressourcen vorliegen, in denen die genannten gesprochensprachlich typischen Phänomene der Lexik konsultiert werden können (vgl. Meliss/Möhrs 2018). Die Ergebnisse der im LeGeDe-Projekt durchgeführten Studien zu der Erwartung künftiger Nutzer an die geplante Ressource zeigen, dass sowohl bei den L1-SprecherInnen als auch bei den L2-SprecherInnen des Deutschen zu jeweils über 70 % Bedarf an einem WB zu Spezifika des gesprochenen Deutsch vorhanden ist. Diese Beobachtung bestätigt die grundsätzliche Annahme zum ansteigenden Bedarf. (iv) Es existieren zurzeit kaum korpusbasierte lexikografische Projekte zur Lexik der gesprochenen Sprache. Lediglich zum Dänischen wurde ein kleines Projekt zu Interjektionen (vgl. Hansen/Hansen 2012) durchgeführt und es liegt ein Paper zum gesprochenen Slowenisch in Verbindung mit lexikografischen Überlegungen vor (vgl. Verdonik/Sepesy Maučec 2017).

Diese vier Grundannahmen sind Ausgangspunkte für die konzeptionellen Überlegungen der geplanten lexikografischen Ressource im LeGeDe-Projekt. Daneben wurden gleich zu Beginn des Projekts die zwei oben bereits erwähnten empirischen Studien zu Erwartungshaltungen an eine lexikografische Ressource mittels zweier Befragungen[7] durchgeführt. Sie fokussieren neben Informationen zu soziodemografischen Daten und zur Nutzung von Online-Ressourcen allgemein hauptsächlich Fragen zur Einschätzung der Relevanz der Spezifik der gesprochenen Lexik in unterschiedlichen Situationen sowie zu Vorstellungen und Erwartungen an eine zukünftige Ressource zur Lexik des gesprochenen Deutsch.

## 3. Gegenstandsbereich und Korpusgrundlage

Der Gegenstandsbereich, mit dem sich das LeGeDe-Projekt beschäftigt, ist die Lexik des gesprochenen Deutsch in unterschiedlichen Interaktionstypen. Die im Fokus stehende Lexik ist durch die Merkmale „standardnah" und „distinktiv" ausgezeichnet, die eine Abgrenzung zu anderen Sprachvarietäten erlauben. Regionale, soziolektale, funktionale oder idiolektale Sprachvarietäten werden in den Analysen nicht behandelt. Im Zentrum des Interesses stehen besonders die distinktiven Merkmale im Vergleich zu der Lexik der geschriebenen Standardsprache. Die LeGeDe-Datengrundlage beruht auf dem Forschungs- und Lehrkorpus Gesprochenes Deutsch (= FOLK[8]). FOLK ist das größte Korpus zum gesprochenen Deutsch in der Interaktion und erweist sich in seiner Zusammensetzung als geeignete Datengrundlage für die geplante LeGeDe-Ressource. Für den korpusbasierten Vergleich mit der geschriebenen Sprache werden entsprechende Teilkorpora aus dem Deutschen Referenzkorpus (= DeReKo[9]) genutzt. FOLK ist integriert in das Archiv für Gesprochenes Deutsch und über die Datenbank für Gesprochenes Deutsch (= DGD)[10] abrufbar. Es beinhaltet Gesprächsaufnahmen aus dem deutschsprachigen Raum in unterschiedlichen privaten, institutionellen und öffentlichen Kontexten mit entsprechenden Transkripten und teilweise auch Videoaufnahmen.

[7] Wenn im weiteren Textverlauf auf die beiden Studien referiert wird, so wird für das durchgeführte Experteninterview die Abkürzung „EXPI" und für die Onlineumfrage „OU" verwendet.

[8] FOLK: http://agd.ids-mannheim.de/folk.shtml; vgl. auch Schmidt 2014a und 2016.

[9] DeReKo: http://www.ids-mannheim.de/direktion/kl/projekte/korpora/releases.html; vgl. Kupietz/Schmidt 2015; Institut für Deutsche Sprache 2017.

[10] DGD: http://dgd.ids-mannheim.de; vgl. auch Schmidt 2014b.

Eine der zentralen Forschungs- und Methodikfragen, mit denen sich das LeGeDe-Projekt beschäftigt, steht in Verbindung mit dem Vergleich zur Lexik der geschriebenen Sprache und damit in Bezug zur Erstellung einer Stichwortliste mit Kandidaten, die möglichst typische Phänomene des Gesprochenen aufweisen. Dazu wurde ein frequenzgesteuertes korpusbasiertes Verfahren entwickelt, bei dem die Häufigkeitsklassen der Lemmata aus FOLK und aus DeReKo ermittelt und diese dann in einem direkten Vergleich betrachtet werden (vgl. Meliss et al. 2018). Dieser Häufigkeitsklassenvergleich wurde zunächst intern genutzt, er ist ab September 2018 aber auch über das frei zugängliche Tool „Lexical Explorer“[11] nachvollziehbar. Dieses Tool wurde für die konkreten Ziele im Rahmen der LeGeDe-Projektarbeit entwickelt und ermöglicht u. a., verschiedene quantitative Daten aus FOLK abzufragen und mit Hilfe von Häufigkeitstabellen bezüglich der Wortverteilung über Wortformen, Kookkurrenzen und Metadaten zu erforschen.

Die Top-10 der möglichen Stichwortkandidaten (vgl. in Abb. 1 die Spalte „Lemma“) vermitteln einen ersten Eindruck von denjenigen Kandidaten, die für die Ressource relevant sind. Die hohe Häufigkeitsklassendifferenz bei den Interjektionen (*ah*, *ach*, *oh*), Abtönungspartikeln (*ja*, *halt*), Gesprächspartikeln (*okay*, *ja*, *na*) und Verben (*gucken*, *kriegen*) weist schon mit dieser vor allem quantitativen Perspektive auf Besonderheiten im Gesprochenen hin, die dann im Projekt durch qualitative Studien genauer analysiert werden.

Study corpus vs. DeReKo

Show/hide columns | CSV | Show 10 entries

| Lemma | Study corpus freq | DeReKo freq | Study corpus HK | DeReKo HK | HK Diff | PoS |
|---|---|---|---|---|---|---|
| okay | 7314 | 199942 | 4 | 14 | 10 | NG |
| ah | 4676 | 152521 | 4 | 14 | 10 | NG |
| ach | 4427 | 445211 | 4 | 13 | 9 | NG |
| ja | 73125 | 10506472 | 0 | 8 | 8 | ADV/NG/PTK |
| gucken | 2989 | 375327 | 5 | 13 | 8 | V |
| oh | 3732 | 313095 | 5 | 13 | 8 | NG |
| halt | 7072 | 802658 | 4 | 12 | 8 | ADV/NG/PTK |
| du | 28145 | 6206378 | 2 | 9 | 7 | PRO |
| kriegen | 2249 | 667783 | 5 | 12 | 7 | V |
| na | 3459 | 520673 | 5 | 12 | 7 | NG |
| Filter Lemma | Filter Study corpus | Filter DeReKo Freq | <9 | <15 | >1 | Filter PoS |

Showing 1 to 10 of 307 entries (filtered from 57.666 total entries)

Abb. 1: Screenshot aus dem „Lexical Explorer“: Top-10 Lemmata mit höchster Häufigkeitsklassendifferenz (HK Diff) zwischen FOLK (= Study corpus) und DeReKo[12]

[11] Der „Lexical Explorer“ kann über OWID$^{plus}$ abgerufen werden: http://www.owid.de/lexex/index_de

[12] Die Berechnungen basieren auf dem Stand vom DGD Release 2.10 (23.05.2018).

Eine genauere Untersuchung der erzielten Stichwortkandidaten verweist auf verschiedene Phänomene bzw. Phänomenbereiche, die sich auch durch Rückbezüge zu einschlägigen Studien – als besonders gesprochensprachlich identifizieren lassen (vgl. Fiehler 2016, Schwitalla $^{4}$2012):

(i) Im Verbalbereich sind aus formal-inhaltlicher und interaktionaler Perspektive u. a. Verwendungsformen mit Distributionsbeschränkungen in der Verbalmorphologie [*ich dachte* (Person, Numerus und Tempus), *guck* (Modus: Imperativ), *meinste* (Verschmelzungen bzw. Komplementierungsmuster), *ich kann kein Deutsch* (absoluter Gebrauch von Modalverben)] und spezifische Bedeutungsvarianten, die häufig nur reduzierte morphologische Paradigmen aufweisen (*geht/gehen* mit einer spezifischen Semantik in der 3. Pers. Sg./Pl. z. B. in den Bedeutungen ‚in Bewegung sein', ‚einigermaßen akzeptabel sein' etc.) für die LeGeDe-Ressource von Interesse.

(ii) Auch im Bereich der Entlehnungen – sowohl aus anderen Sprachen (Anglizismen: *cool*, *okay* etc.) als auch aus deutschen Sprachvarietäten (soziolektal: *geil*, *krass*, *Mama*, *wieso* etc.) – zeigen sich interessante, gesprochensprachliche Phänomene, die u. a. in Zusammenhang mit ihrer Gebrauchsfrequenz, dem Vorkommen in bestimmten Interaktionstypen, den Sprechergruppen, der grammatischen Integration und der phonetischen Realisierung stehen.

(iii) Im Bereich der Wortbildung lassen sich u. a. zahlreiche Affixe mit formalen Besonderheiten und/oder neuen Bedeutungsnuancen verbuchen:

*rum-* (*-ärgern, -blödeln, -gammeln, -hängen, -laufen, -machen, -sitzen etc.*),
*rein-* (*-bringen, -gehen, -kommen, -machen, -passen, -tun etc.*),
*rauf-* (*-gehen, -schmieren etc.*),
*mega-* (*-crazy, -einfach, -gut, -krass, -lang, -schlecht, -schwer, -stressig, -viel etc.*),
*super-* (*-cool, -easy, -gut, -lecker, -safe, -schön, -toll, -viel etc.*),
*sau-* (*-blöd, -dumm, -geil, -gut, -teuer etc.*),
*-mäßig* (*alters-, freizeit-, hobby-, kosten-, vereins-, zahlen- etc.*).

(iv) Die Untersuchung zu Gebrauch und Bedeutung von Passepartoutwörtern (machen, tun, Ding, Sache etc.) sowie von Ausdrücken und Formeln der Vagheit (*irgend-* [*-wann, -was, -wie, -wo*] etc.) stehen v. a. in Verbindung mit der Bedeutungskonstituierung in der Interaktion.

(v) Die Verwendung von Teilsynonymen (*werfen/schmeißen*, *kriegen/bekommen/erhalten*, *gucken/schauen/sehen*, *warum/weshalb/wieso*, *darum/deshalb/deswegen* etc.) interessiert insbesondere aus der Perspektive von Stil und Register.

(vi) Die für die gesprochene Sprache besonders typischen Interjektionen (*ach, ah, oh* etc.) sowie Diskurspartikeln (*einfach*, *genau*, *gut*, *okay*, *richtig*, *sicher* etc.) und Abtönungspartikeln (*eben*, *halt*, *mal* etc.) bilden ebenfalls einen interessanten und umfangreichen Bereich des anvisierten Gegenstandsbereiches.

(vii) Schließlich sind von ganz zentralem Interesse auch zahlreiche Verwendungsweisen bzw. feste Verwendungsmuster in Verbindung mit einer Verbalform (*guck mal*, *wenn du meinst, sag ich mal*, *wenn's sein muss*, *was weiß ich*, *ich weiß nicht* etc.), einer Nominalform (*keine Ahnung*, *mein Gott*, *Gott sei Dank*, *in Ordnung* etc.) oder adjektivischen bzw. adverbialen Elementen (*alles klar*, *na gut*, *na schön* etc.), die im interaktionalen Kontext eine spezifische Funktion besitzen.

Der Gegenstandsbereich wurde auch in den durchgeführten empirischen Befragungen thematisiert. Diesbezüglich kann daher an die Ergebnisse der Frage „*Welche Art von Stichwörtern würden Sie in einem WB des gesprochenen Deutsch erwarten?*" aus der OU angeknüpft werden, die besonders den Wunsch nach Stichwörtern, die eine spezifische Kombinatorik aufweisen, deutlich unterstreichen. Jeweils rund drei Viertel aller Befragten der OU erwarten in einem WB des gesprochenen Deutsch sowohl Stichwörter, die eine formelhafte Verwendung (79,5 %) besitzen, als auch Stichwörter mit einem – im Vergleich zur geschriebenen Sprache – anderen Kombinationspotenzial (74,6 %). Der Vergleich mit entsprechenden Antworten zu einer parallelen Frage aus dem EXPI zeigt ähnliche Ergebnisse. Lexikalische Einheiten, die in der gesprochenen Sprache ein anderes Kombinationspotenzial als in der geschriebenen Sprache aufweisen, stehen bei den ExpertInnen an oberster Stelle auf der Wunschliste von möglichen Stichwörtern (94,1 %). Dies beinhaltet – laut Äußerungen der interviewten ExpertInnen – Konstruktionen, lexikalische Ausdrücke, syntagmatische Kombinationen, Formeln, Chunks etc. und auch Mehrwortlemmata.

## 4. Datengewinnung und lexikografische Umsetzung

Die einzelnen Phasen und Arbeitsschritte zur Entwicklung der LeGeDe-Ressource folgen grundsätzlich dem lexikografischen Prozess zur Erstellung von Internetwörterbüchern, wie es in Klosa/Tiberius (2016: 75f.) beschrieben wird (vgl. dazu einige exemplarisch ausgewählte Daten zu Pilotstudien in Abschnitt 4.1). Die lexikografische Umsetzung stellt sich dabei neuen Herausforderungen, die in Verbindung mit einer Reihe von unterschiedlichen lexikografischen Fragestellungen stehen: Welche Stichwortansetzung kann sinnvollerweise für ein Wörterbuch zur gesprochenen Sprache verfolgt werden (z. B. Einwortlemmata, Mehrwortlemmata)? Welche Benutzungssituation wäre für ein Wörterbuch zum Gesprochenen denkbar? Wie können authentische mündliche Sprachdaten (z. B. Audiodateien, Transkripte) in die Ressource integriert werden? Wie können Informationen zu Form, Inhalt und Funktion von gesprochensprachlichen lexikalischen Elementen in der Mikrostruktur der geplanten Ressource (vgl. hierzu Abschnitt 4.2) angemessen dargestellt werden?

### 4.1. Gewinnung, Kodierung, Analyse und Strukturierung der Daten

Verschiedene Verfahren zur Bedeutungsdisambiguierung und zur korpusbasierten Entwicklung von Wortprofilen, die als methodologische Ansätze in der korpusbasierten Lexikologie und Lexikografie gelten, sollen mit der Analyse sprachlicher Einheiten und der Beschreibung von kommunikativen Funktionen aus Sicht der Interaktionslinguistik (vgl. u. a. Deppermann 2007) vereint werden. Unterschiedliche Informationen, wie u. a. automatisch generierte Daten (Frequenzdaten zum Formenbestand, zu Kollokationen, Kookkurrenzen etc.) sollen nicht nur bei der ersten korpusbasierten Annäherung an die Daten durch entsprechende Hypothesenbildung, sondern auch bei der Interpretation der Analyseergebnisse unterstützen. In ersten Studien wurden die Gesprächssequenzen zu entsprechenden Belegen nach Form, Bedeutung und Funktion in Verbindung mit metasprachlicher Information analysiert, die Ergebnisse strukturiert und hinsichtlich der Relevanz der Ressource interpretiert. Datengrundlage für die Analyse ist jeweils eine Zufallsstichprobe aus FOLK. Zu den Korpusbelegen stehen in allen Fällen Audioaufnahmen zur Verfügung, die über die Datenbank parallel zu den Transkripten abgerufen werden können. Neben Metadaten zum Treffer und zum Transkript, die automatisch extrahiert vorlie-

gen, können in manuellen Untersuchungen inhaltlich funktionale, syntaktisch-formale, sequenzrelevante sowie grammatische Aspekte im Abgleich mit den Metadaten betrachtet werden. Die Möglichkeiten dieses Verfahrens lassen sich exemplarisch an dem Verballemma *gucken* aufzeigen. Die erzielten Kodierungs- und Strukturierungsergebnisse von 100 Belegen einer Zufallsstichprobe ergaben für *gucken* zunächst mindestens neun verschiedene Bedeutungsvarianten, von denen nur zwei im LGWB-DaF und dem DWDS sowie drei in Duden-online verzeichnet sind. Die semantische Disambiguierung erfolgte durch ein Zusammenspiel von Kodierungen bezüglich der Form (Strukturmuster) und der spezifischen Bedeutungen (vgl. dazu Auszüge aus dem Bedeutungsspektrum des Lemmas *gucken* auf Basis der LeGeDe-Stichprobe in Meliss/Möhrs 2017: 47, Möhrs/Meliss/Batinić 2017: 293). Zusätzlich zu diesen Beobachtungen, die die inhaltliche Beschreibungsebene betreffen, finden sich in der Stichprobe spezielle, distributionell eingeschränkte und/oder modalisierende Verwendungsformen (Imperativ / Modalpartikel / Modalverb: *guck (mal); (muss) mal gucken*). Eine auf diesen Beobachtungen aufbauende Analyse und Systematisierung der Ergebnisse zeigen weitere musterhafte Verwendungsmöglichkeiten von *gucken* mit unterschiedlichen interaktionalen Funktionalitäten, z. B. als Diskursmarker auf (vgl. dazu u. a. Günthner 2017, Wegner 2015, Proske 2017).

### 4.2. Überlegungen zur Mikrostruktur der geplanten Ressource

Erste Überlegungen zur Entwicklung einer lexikografischen Mikrostruktur der LeGeDe-Ressource verfolgen ein breit gefächertes Informationsangebot. Jeder Wortartikel enthält bedeutungs- und funktionsübergreifende frequenzorientierte Informationen, die (semi-) automatisch und teilweise auch im Vergleich zu Daten aus DeReKo generiert werden, zum Beispiel zu Häufigkeiten und Häufigkeitsklassen. Weitere frequenzorientierte Angaben zum Formeninventar (vgl. Tab. 1), zur Kombinatorik (Bigramme, Trigramme und Kookkurrenzen; vgl. zu Kookkurrenzen Abb. 2) oder zu Metadaten (vgl. Abb. 3) können über den „Lexical Explorer" abgerufen werden und sind Ausgangspunkt für weitere qualitative lexikografische Analysen. Ein Link aus dem Wortartikel zum Tool ermöglicht den NutzerInnen weitere explorative Untersuchungen. Entsprechende korpuslinguistische Werkzeuge (wie zum Beispiel eine Kookkurrenzanalyse) sind in bisherigen lexikografischen Projekten (vgl. z. B. das DWDS oder *elexiko*) zur automatischen Analyse und Strukturierung von Sprachdaten nicht neu. Allerdings wurden diese Verfahren bislang noch nicht auf Gesprächsdaten angewendet. Fehlende Satzgrenzen, Überlappungen, Pausen oder Abbrüche sind nur einige der Herausforderungen, die bei der Erarbeitung von diesen Werkzeugen auch für gesprochensprachliches Material zu bedenken sind. Frequenzorientierte Korpusabfragen sowie die Abfragen quantitativer Daten über den „Lexical Explorer" werden im Folgenden an verschiedenen Beispiellemmata vorgestellt.

(i): Durch frequenzbasierte Daten zum Formeninventar in FOLK und DeReKo, die hier durch das Häufigkeitsklassenmaß (Frequenzklassen) und das entsprechende Maß für Frequenzklassendifferenzen angegeben werden, lassen sich u. a. Schlüsse zu gesprochensprachlichen Besonderheiten ziehen. So geben beispielsweise die Daten zur Frequenzklasse für *denke* (z. B. 1. Pers. Sg. Indikativ Aktiv Präsens) mit der HK-Klasse „6" in FOLK und einer Häufigkeitsklassendifferenz zu Gunsten von FOLK im Vergleich zu DeReKo (HK-Differenz) von „3" (vgl. Tab. 1) einen Hinweis darauf, dass die Verbindung *ich denke* besondere gesprochensprachliche Funktionen haben kann (z. B. als „gesprächsstrukturelles Gliederungssignal", vgl. Zeschel 2017: 291).

| Normalisierte Form | POS | FOLK Freq.-Klasse | DeReKo Freq.-Klasse | Differenz |
|---|---|---|---|---|
| **denkst** | VVFIN | 9 | 14 | 5 |
| **denk** | VVIMP | 10 | 14 | 4 |
| **denken** | VVFIN | 9 | 13 | 4 |
| **dachte** | VVFIN | 7 | 10 | 3 |
| **denke** | VVFIN | 6 | 9 | 3 |
| **denkt** | VVFIN | 9 | 10 | 1 |

Tab. 1: Automatisch generierte Information zu Frequenzklassen im Vergleich FOLK und DeReKo in Verbindung mit dem Formeninventar zu dem Lemma *denken*[13]

(ii) Automatisch generierte Angaben zur Kombinatorik in Form von Bi- oder Trigrammen und/oder Kookkurrenzprofilen bieten Information zu häufigen Kollokationspartnern. So ist z. B. *gut* als Kookkurrenzpartner von *okay* auffällig frequent in FOLK (LLR: 989.96, ABS: 232[14]; vgl. Abb. 2). Die Information zu Kollokationsmöglichkeiten (*okay gut*; *gut okay*) in Verbindung mit der Position der Partner in der Gesprächssequenz weisen an initialer Position bzw. finaler Position auf jeweils unterschiedliche interaktionsspezifische Funktionalitäten hin. Diese müssen durch weitere ausführliche Sequenzanalysen erfasst und für die lexikografische Umsetzung redaktionell ausgearbeitet werden.

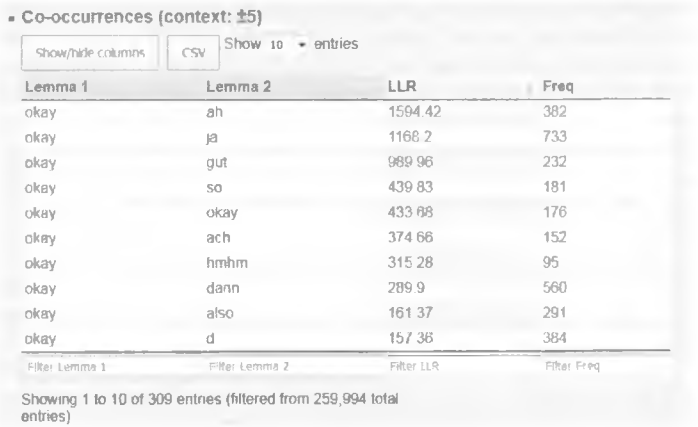
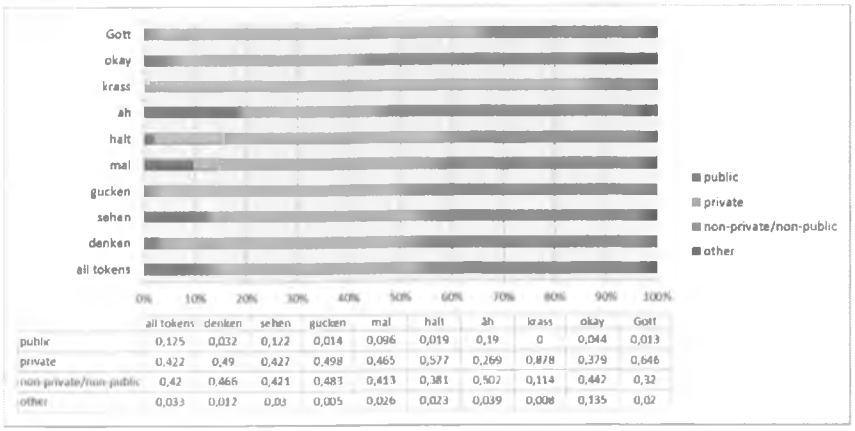



Abb. 2: Screenshot aus dem „Lexical Explorer": Kookkurrenzen zum Lemma *okay*

(iii) Die umfangreichen Metadaten, die zu FOLK vorliegen, erlauben unter anderem, die Verteilung auf unterschiedliche Interaktionstypen zu untersuchen und ggf. auch Vergleiche zwischen Stichwörtern ähnlicher Klassen (*gucken/sehen*; *halt/mal*) zu ziehen (vgl. Abb. 3).

[13] Die Berechnungen basieren auf dem Stand vom DGD Release 2.8 (27.11.2017).

[14] LLR = Log-likelihood ratio; ABS = absolute Häufigkeit. Die Berechnungen basieren auf dem Stand vom DGD Release 2.10 (23.05.2018).




| | all tokens | denken | sehen | gucken | mal | halt | äh | krass | okay | Gott |
|---|---|---|---|---|---|---|---|---|---|---|
| public | 0,175 | 0,032 | 0,122 | 0,014 | 0,096 | 0,019 | 0,19 | 0 | 0,044 | 0,013 |
| private | 0,422 | 0,49 | 0,427 | 0,498 | 0,465 | 0,577 | 0,269 | 0,878 | 0,379 | 0,646 |
| non-private/non-public | 0,42 | 0,466 | 0,421 | 0,483 | 0,413 | 0,381 | 0,502 | 0,114 | 0,442 | 0,32 |
| other | 0,033 | 0,012 | 0,03 | 0,005 | 0,026 | 0,023 | 0,039 | 0,008 | 0,135 | 0,02 |

Abb. 3: Verteilung unterschiedlicher lexikalischer Einheiten auf Interaktionstypen (im Vergleich zur Verteilung aller Einheiten; siehe „all tokens")[15]

Die Verteilung auf Interaktionstypen im Vergleich zur Anzahl der Gesamttokens zeigt z. B., dass *krass* in FOLK kein Vorkommen in öffentlichen Interaktionen, aber dafür ein sehr hohes Aufkommen in privaten Interaktionskontexten aufweist.

Das weitere Informationsangebot besteht aus verschiedenen Arten von redaktionell erarbeiteten lexikologischen Informationen. Diese umfassen z. B. Angaben zu Auffälligkeiten und Unterschieden bezüglich Form und Bedeutung. Einheiten mit interaktionaler Funktion werden zudem anhand von interaktionstypischen Parametern beschrieben (z. B. Funktion in der Interaktion, Syntax und sequenzielle Realisierung, Prosodie).

Auch in der OU wurde nach den Wünschen zum Informationsangebot gefragt: „*Welche Informationen sollten, Ihrer Meinung nach, in einem Wörterbuch des gesprochenen Deutsch angeboten werden?*" (vgl. Abb. 4).

[15] Einige Beispiele für die unterschiedlichen Kategorien: „*public*": Schlichtungsgespräch, Podiumsdiskussion etc., „*private*": Tischgespräch, Paargespräch, Telefongespräch, Gespräch beim Kochen etc., „*non private/non public*": Lehr- und Lernkontexte, berufliche Kommunikation, Verkaufsgespräche etc. und „*other*": Maptask. Die Berechnungen basieren auf dem Stand vom DGD Release 2.8 (27.11.2017).

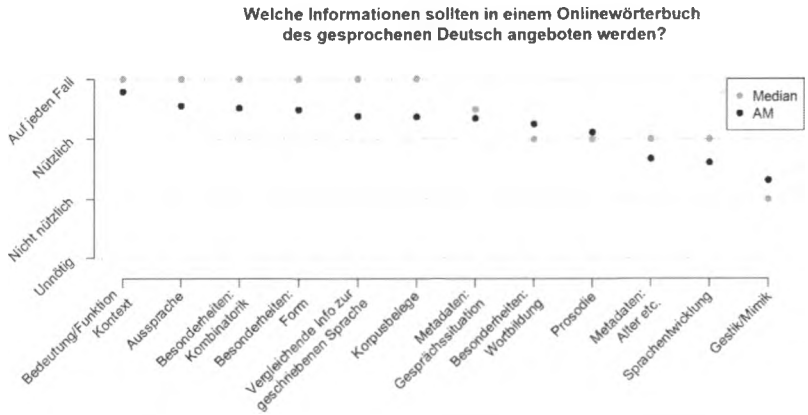



Abb. 4: Ergebnisse der OU zur Frage „Welche Informationen sollten, Ihrer Meinung nach, in einem Wörterbuch des gesprochenen Deutsch angeboten werden?"

Aus den 5 verschiedenen Antwortoptionen von „*Auf jeden Fall*" (Rang 1) bis hin zu „*Ich weiß nicht*" (Rang 5) ergibt sich eine Rangfolge, in der die Ränge 1-3 entsprechende Antworten aufweisen. Die Ergebnisse der OU – in Abb. 4 durch den durchschnittlichen Mittelwert visualisiert – zeigen, dass u. a. Informationen zu Besonderheiten in der Kombinatorik mit den Antworten „*Auf jeden Fall*" bzw. „*Nützlich, jedoch nicht zwingend erforderlich*" von den Befragten zwischen Platz 1 und 2 bewertet wurden. Wenn man die Ergebnisse aus dem EXPI zum Vergleich heranzieht, dann wird deutlich, dass die Informationen zu Aussprache, Bedeutung im Kontext, Besonderheiten in der Form, Besonderheiten in der Kombinatorik, Angebot von Korpusbelegen, Metadaten zur Gesprächssituation und Prosodie gleichermaßen an oberster Stelle mit „*Auf jeden Fall*" gewünscht wurden (vgl. Meliss/Möhrs/Ribeiro Silveira 2018: 126f.).

Zu den verschiedenen Angabetypen soll es in der geplanten LeGeDe-Ressource Möglichkeiten für lexikografische Kommentare und Hinweise geben, die durch ausgewählte Belege aus FOLK illustriert werden. Diese Korpusbelege bestehen aus einem von LexikografInnen formulierten Belegtitel, einer Belegkontextinformation, dem Transkriptausschnitt mit entsprechender Verlinkung zur Audiodatei sowie einem optionalen Textbaustein, in dem der Belegausschnitt genauer analysiert wird. Insgesamt ist vorgesehen, verschiedene Formen und Wege des Zugriffs zu ermöglichen. Diesbezüglich wurde auch die Frage „*Welche Zugriffsmöglichkeiten halten Sie für relevant?*" der schon erwähnten OU recht einstimmig beantwortet (vgl. Meliss/Möhrs/Ribeiro Silveira 2018: 128f.). Alle Befragten der OU halten eine alphabetische Anordnung, einen Zugriff über eine Suchmaske (erweiterte Suche), einen Zugriff über kommunikative Funktionen und über Themen- und Bedeutungsfelder für besonders relevant (zwischen Rang 1: „*Auf jeden Fall*" und Rang 2: „*Nützlich, jedoch nicht zwingend erforderlich*").

## 5. Ausblick

Das bis 2019 erarbeitete Produkt[16], das als eine Art Prototyp für ein korpusbasiertes Wörterbuch zur Lexik des gesprochenen Deutsch angedacht ist, wird in die Plattform OWID$^{plus}$[17] am IDS eingebunden. Während der Projektarbeit standen vor allem die konzeptionellen Überlegungen im Mittelpunkt mit dem Ziel, zu ausgewählten Stichwortkandidaten prototypische Einträge zu erarbeiten. Mit der Umsetzung des LeGeDe-Wörterbuchs als Prototyp soll eine Möglichkeit präsentiert werden, wie Besonderheiten der Lexik des gesprochenen Deutsch auf der Basis von authentischen gesprochensprachlichen Daten lexikografisch zugänglich gemacht werden können. Zahlreiche empirisch erhobene Erwartungen von zukünftigen Nutzern konnten in die Umsetzung aufgenommen werden.

## Literatur


Deppermann, Arnulf (2007): *Grammatik und Semantik aus gesprächsanalytischer Sicht*. Berlin: de Gruyter.

Deppermann, Arnulf/Proske, Nadine/Zeschel, Arne (Hg.) (2017): *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch*. Tübingen: Narr.

Fiehler, Reinhard (2016): „Gesprochene Sprache". In: Wöllstein, Angelika (Hg.): *Duden – Die Grammatik*. Berlin: Dudenverlag, S. 1181-1260.

Günthner, Susanne (2017): „Diskursmarker in der Interaktion – Formen und Funktionen univerbierter guck mal- und weißt du-Konstruktionen". In: Blühdorn, Hardarik/Deppermann, Arnulf/Helmer, Henrike/Spranz-Fogasy, Thomas (Hg.) (2017): *Diskursmarker im Deutschen. Reflexionen und Analysen*. Göttingen: Verlag für Gesprächsforschung, S. 103-130.

Handwerker, Brigitte/Bäuerle, Rainer/Sieberg, Bernd (Hg.) (2016): *Gesprochene Fremdsprache Deutsch* [= Perspektiven Deutsch als Fremdsprache, Band 32]. Baltmannsweiler: Hohengehren.

Hansen, Carsten/Hansen, Martin H. (2012): „A Dictionary of Spoken Danish". In: Fjeld, Ruth Vatvedt/Torjusen, Julie Matilde (Hg.): *Proceedings of the 15th EURALEX International Congress*. 7-11 August 2012. Oslo, Norway: Department of Linguistics and Scandinavian Studies, University of Oslo, S. 929-935.

Imo, Wolfgang (2007): *Construction Grammar und Gesprochene-Sprache-Forschung. Konstruktionen mit zehn matrixsatzfähigen Verben im gesprochenen Deutsch* [= Germanistische Linguistik, Band 275]. Tübingen: Niemeyer.

Imo, Wolfgang/Moraldo, Sandro M. (Hg.) (2015): *Interaktionale Sprache und ihre Didaktisierung im DaF-Unterricht* [= Deutschdidaktik, Band 4]. Tübingen: Stauffenburg.

Institut für Deutsche Sprache (2017): *Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2017-I* (Release vom 08.03.2017). Mannheim: Institut für Deutsche Sprache. In: http://www.ids-mannheim.de/direktion/kl/projekte/korpora/releases.html (letzter Zugriff: 06.08.2019).

Klosa, Annette/Tiberius, Carole (2016): „Der lexikografische Prozess". In: Klosa, Annette/Müller-Spitzer, Carolin (Hg.): *Internetlexikografie. Ein Kompendium*. Unter Mitarbeit von Martin Loder. Berlin/Boston: de Gruyter, S. 65-110.

Kupietz, Marc/Schmidt, Thomas (2015): „Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung". In: Eichinger, Ludwig M. (Hg.): *Sprachwissenschaft im Fokus* [= Jahrbuch des Instituts für Deutsche Sprache 2015]. Berlin/Boston: de Gruyter, S. 297-322.

LGWB-DaF = *Langenscheidt Großwörterbuch Deutsch als Fremdsprache* (Neubearbeitung 2015).


[16] Die Freischaltung der LeGeDe-Ressource ist für die 2. Jahreshälfte 2019 geplant (www.owid.de/legede/).

[17] OWID$^{plus}$: https://www.owid.de/plus/index.html.

Meliss, Meike (2016): „Gesprochene Sprache in DaF-Lernerwörterbüchern". In: Handwerker, Brigitte et al. (Hg.): *Gesprochene Fremdsprache Deutsch* [= Perspektiven Deutsch als Fremdsprache, Bd. 32]. Baltmannsweiler: Hohengehren, S. 179-199.

Meliss, Meike/Möhrs, Christine (2017): „Die Entwicklung einer lexikografischen Ressource im Rahmen des Projektes LeGeDe". In: *Sprachreport* 4/2017, S. 42-52.

Meliss, Meike/Möhrs, Christine (2018): „Lexik in der spontanen, gesprochensprachlichen Interaktion: Eine Annäherung aus der DaF-Perspektive". In: *GFL-Journal* [= German as a foreign language] 3/2018, S. 79-110.

Meliss, Meike/Möhrs, Christine/Ribeiro Silveira, Maria (2018): „Erwartungen an eine korpusbasierte lexikografische Ressource zur 'Lexik des gesprochenen Deutsch in der Interaktion': Ergebnisse aus zwei empirischen Studien". In: *Zeitschrift für Angewandte Linguistik*. 1/2018, S. 103-138.

Meliss, Meike/Möhrs, Christine/Ribeiro Silveira, Maria (2019): „Anforderungen und Erwartungen an eine lexikografische Ressource des gesprochenen Deutsch aus der L2-Lernerperspektive". In: *Lexicographica* 34(1), S. 89-121.

Meliss, Meike/Möhrs, Christine/Batinić, Dolores/Perkuhn, Rainer (2018): „Creating a List of Headwords for a Lexical Resource of Spoken German". In: Čibej, Jaka/Gorjanc, Vojko/Kosem, Iztok/Krek, Simon (Hg.): *Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts*, 17-21 July, Ljubljana. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, S. 1009-1016.

Möhrs, Christine/Meliss, Meike/Batinić, Dolores (2017): „LeGeDe - towards a corpus-based lexical resource of spoken German". In: Kosem, Iztok/Tiberius, Carole/Jakubíček, Miloš/Kallas, Jelena/Krek, Simon/Baisa, Vít (Hg.): *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Leiden, the Netherlands, 19.-21. September 2017. Brno: Lexical Computing CZ s.r.o., 2017, S. 281-298.

Moon, Rosamund (1998): „On using spoken data in corpus lexicography". In: Fontenelle, Thierry/Hiligsmann, Philippe/Michiels, Archibald/Moulin, André/Theissen, Siegfried (Hg.) (1998): *Proceedings of the 8th EURALEX International Congress*. 4-8 August 1998. Liège, Belgium: English and Dutch Departments, University of Liège, S. 347-355.

Moraldo, Sandro M./Missaglia, Federica (Hg.) (2013): *Gesprochene Sprache im DaF-Unterricht. Grundlagen – Ansätze – Praxis* [= Sprache – Literatur und Geschichte, Band 43]. Heidelberg: Winter.

Proske, Nadine (2017): „Zur Funktion und Klassifikation gesprächsorganisatorischer Imperative". In: Blühdorn, Hardarik/Deppermann, Arnulf/Helmer, Henrike/Spranz-Fogasy, Thomas (Hg.) (2017): *Diskursmarker im Deutschen. Reflexionen und Analysen*. Göttingen: Verlag für Gesprächsforschung, S. 73-102.

Reeg, Ulrike/Gallo, Pasquale/Moraldo, Sandro M. (Hg.) (2012): *Gesprochene Sprache im DaF-Unterricht. Zur Theorie und Praxis eines Lerngegenstandes* [= Interkulturelle Perspektiven in der Sprachwissenschaft und ihrer Didaktik, Band 3]. Berlin: Waxmann

Schmidt, Thomas (2014a): „The Research and Teaching Corpus of Spoken German – FOLK". In: *Proceedings of LREC'14*, Reykjavik, Iceland: ELRA.

Schmidt, Thomas (2014b): „The Database for Spoken German - DGD2". In: *Proceedings of LREC'14*, Reykjavik, Iceland: ELRA.

Schmidt, Thomas (2016): „Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German". In: Kirk, John M./Andersen, Gisle (Hg.): *Compilation, transcription, markup and annotation of spoken corpora, Special Issue of the International Journal of Corpus Linguistics* [IJCL 21:3], S. 396-418.

Schwitalla, Johannes (2012): *Gesprochenes Deutsch. Eine Einführung*. 4., neu bearbeitete und erweiterte Auflage [= Grundlagen der Germanistik 33]. Berlin: Schmidt.

Sieberg, Bernd (2013): *Sprechen lehren, lernen und verstehen. Stufenübergreifendes Studien- und Übungsbuch für den DaF-Bereich*. Tübingen: Julius Groos.

Trap-Jensen, Lars (2004): „Spoken Language in Dictionaries: Does it Really Matter?". In: Williams, Geoffrey/Vessier, Sandra (Hg.) (2004): *Proceedings of the 11th EURALEX International Congress*. 6-10 July 2004. Lorient, France: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud, S. 311-318.

Trim, John/North, Brian/Coste, Daniel/Sheils, Joseph (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen. Niveau A1, A2, B1, B2, C1, C2*. Übersetzt von Jürgen Quetz, Raimund Schieß, Ulrike Sköries und Günther Schneider. Europarat. Rat für kulturelle Zusammenarbeit. Herausgegeben vom Goethe-Institut Inter Nationes, der Ständigen Konferenz der Kulturminister der Länder in der Bundesrepublik Deutschland (KMK), der Schweizerischen Konferenz der Kantonalen Erziehungsdirektoren (EDK) und dem österreichischen Bundesministerium für Bildung, Wissenschaft und Kulture (BMBWK). Berlin [u. a.]: Langenscheidt.

Verdonik, Darinka/Sepesy Maučec, Mirjam (2017): „A Speech Corpus as a Source of Lexical Information". In: *International Journal of Lexicography* 30/2, S. 143-166.

Wegner, Lars (2015): „„...mal kucken/schauen/sehen..."-Konstruktionen in Elternsprechtagsgesprächen – zur engen Verknüpfung von syntaktischen Konstruktionen und kommunikativen Gattungen". In: Bücker, Jörg/Günthner, Susanne/Imo, Wolfgang (Hg.) (2015): *Konstruktionsgrammatik V. Konstruktionen im Spannungsfeld von sequenziellen Mustern, kommunikativen Gattungen und Textsorten* [= Stauffenburg Linguistik, Band 77]. Tübingen: Stauffenburg, S. 163-186.

Zeschel, Arne (2017): „Denken und wissen im gesprochenen Deutsch". In: Deppermann, Arnulf/Proske, Nadine/Zeschel, Arne (Hg.) (2017): *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch* [= Studien zur deutschen Sprache, Band 74] Tübingen: Narr, S. 249-336.

## Online-Ressourcen

DEREKO = Deutsches Referenzkorpus, http://www.ids-mannheim.de/direktion/kl/projekte/korpora/releases.html (letzter Zugriff: 06.08.2019).

DGD = Datenbank für Gesprochenes Deutsch, http://dgd.ids-mannheim.de (letzter Zugriff: 06.08.2019).

Duden-online, https://www.duden.de/ (letzter Zugriff: 06.08.2019).

DWDS = Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart, www.dwds.de (letzter Zugriff: 06.08.2019).

FOLK = Forschungs- und Lehrkorpus Gesprochenes Deutsch, http://agd.ids-mannheim.de/folk.shtml (letzter Zugriff: 06.08.2019).

LeGeDe (Projektwebseiten), http://www.ids-mannheim.de/lexik/lexik-des-gesprochenen-deutsch.html (letzter Zugriff: 06.08.2019); LeGeDe-Ressource, http://www.owid.de/legede/ (Freischaltung in Kürze).

Lexical Explorer, http://www.owid.de/lexex/index_de (letzter Zugriff: 06.08.2019).

OWID = Online Wortschatz Informationssystem Deutsch, www.owid.de (letzter Zugriff: 06.08.2019).