Creating a List of Headwords for a Lexical Resource of Spoken German

Meike Meliss, Christine Möhrs, Dolores Batinić, Rainer Perkuhn
Institut für Deutsche Sprache, Mannheim
E-mail: meliss@ids-mannheim.de, moehrs@ids-mannheim.de, batinic@ids-mannheim.de, perkuhn@ids-mannheim.de

Abstract

Except for some recent advances in spoken language lexicography (cf. Verdonik & Sepesy Maučec 2017, Hansen & Hansen 2012, Siepmann 2015), traditional lexicographic work is mainly oriented towards the written language. In this paper, we describe a method we used to identify relevant headword candidates for a lexicographic resource for spoken language that is currently being developed at the Institute for the German Language (IDS, Mannheim). We describe the challenges of the headword selection for a dictionary of spoken language, and having made considerations regarding our headword concept, we present the corpus-based procedures that we used in order to facilitate the headword selection. After presenting the results regarding the selection of one-word lemmas, we discuss the opportunities and limitations of our approach.

Keywords: list of headwords, spoken German, corpus-based methods

1 Introduction

In the project "Lexik des gesprochenen Deutsch" (=LeGeDe)[1] a corpus-based lexical resource for standard spoken German in interaction is being developed. In this resource, lexical and interaction-specific features are to be gathered and presented in a multimedia manner (cf. Meliss & Möhrs 2017; Möhrs, Meliss & Batinić 2017). Identifying and defining the characteristics of spoken German are important tasks for creating this new type of lexicographic resource. These tasks are directly related to the selection of headwords and to the methodological procedures for creating the resource (see Section 4).

The results of two surveys on the expectations for and requirements of a resource for the standard lexicon of spoken German, (cf. Meliss, Möhrs & Ribeiro Silveira 2018) confirm that the lexicographic codification of spoken language and its interactional characteristics are not satisfactorily addressed in current dictionaries (cf. Meliss 2016: 195; Eichinger 2017: 283), despite the call made almost 15 years ago by Trap-Jensen that "we should pay more attention to spoken language when making our dictionaries" (2004: 311). Apart from particular recent advances in spoken language lexicography (cf. Verdonik & Sepesy Maučec 2017; Hansen & Hansen 2012; Siepmann 2015), both consideration of and experience with spoken language in lexicography are still rare. Hence, the LeGeDe resource can barely rely on existing models that could provide guidance in the creation of the list of headwords.

For developing this novel type of a spoken language lexical resource, the definition of "headword" from the research tradition of dictionaries based on written language has to be supplemented with new perspectives to address the peculiarities of the spoken form (cf. Deppermann, Proske & Zeschel

[1] The project LeGeDe (http://www.ids-mannheim.de/lexik/lexik-des-gesprochenen-deutsch.html; [16/05/2018]) is a third-party funded project of the Leibniz Association. It is being implemented as a cooperation project of the Department of Pragmatics and Lexical Studies at the Institute for the German Language in Mannheim and has a duration of three years (2016-2019); for a detailed project description, see Meliss & Möhrs 2017.

2017). The headword concept has to include considerations that take into account the one-word lemmas that have specific meanings and uses in spoken language in comparison to written language. In addition to one-word lemmas, multi-word expressions and constructions that have specific functions in interaction (e.g. *ich weiß nicht* or *keine Ahnung*; cf. Bergmann 2017) are also of interest as headword candidates (see Section 5).

The aim of our contribution is to present the corpus-based and interpretative method we used for detecting salient terms of typical spoken lexicon. In this study we focus on one-word lemmas (see Sections 3 and 4).

2 The Project LeGeDe

The aim of the LeGeDe project is to develop a corpus-based electronic resource that addresses lexical peculiarities of spoken German in interaction. The specifics of spoken language lexis were rather neglected in previous lexicographic codifications and research (Meliss & Möhrs 2017: 47). Within the framework of the (present) project work, however, they are to be identified, analyzed and described via different corpus-based and interpretative methods.

The subject matter covered by the LeGeDe project is the lexicon of spoken German characterized by the feature “standard”[2], which allows a differentiation to other medial language varieties. Of particular interest are the distinctive features of the spoken lexicon in comparison to the lexicon of the written standard language. In order to find these, we work with the largest corpus of spoken German in interactional settings (FOLK: Research and Teaching Corpus of Spoken German, Schmidt 2014; 1,96 Million tokens). One of the key research and methodology issues addressed by the LeGeDe project is the detection of differences of the lexicon of written and spoken language, and thus the selection and description of headwords that represent the most typical and distinctive phenomena of spoken lexis.

3 Corpus-based procedures for assisting the creation of a list of headwords

In order to assist the selection of headword candidates for the LeGeDe resource, we performed a lemma comparison between FOLK and the German reference corpus of written German (DeReKo 2017 I, cf. Kupietz/Keibel 2009; 30 Billion tokens). Since we used DeReKo as a representation of current written language, we excluded the data containing conceptual spoken language represented in Wikipedia discussions, as well as the sub-corpus “Sprachliche Umbrüche”, dating from 1945 to 1968. FOLK is lemmatized automatically with TreeTagger (Schmid 1994) by using a parameter file trained on a manually annotated gold standard (Westpfahl & Schmidt 2016). For DeReKo several different annotation layers are provided (with many different suggestions about lemmatization, lemma forms, and, also, part-of-speech information, cf. Stadler 2014). For ease of comparison of spoken and written data, we selected only the highest ranked suggestion of the TreeTagger lemmatization. The TreeTagger was applied to the written data with an especially customized parameter file that was prepared by the author of the tool for DeReKo.

Firstly, we simplified the part-of-speech tags in FOLK (cf. Westpfahl 2014) to more universal categories (V for verbs, N for nouns etc.) to facilitate the headword selection. Then, we aggregated the part-of-speech tags with which a lemma has been tagged in FOLK (V/N if a word was tagged as verb and as a noun). In addition, we marked the lemmas that we did not want to consider in further examinations as outliers. These were: lemmas tagged as proper names, numerals, lemmas affected by orthographic reforms, idiosyncrasies and lemmas that occurred only in one transcript.

[2] This means that the project will not consider dialects (such as Bavarian), sociolects (such as adolescent language) or idiolects.

We calculated the difference in lemma distribution in the corpora by using different effect size measures (odds ratio, %diff, relative risk, binary log of relative risk, and frequency classes) and measures of statistical significance (log likelihood ratio and chi square). Then, we integrated the table containing the lemma comparison into a tool we developed for facilitating, filtering and sorting the data in a fast and user-friendly way. With the help of this tool, the headword candidates can be assessed, performed and explored dynamically, and the parameters can be adjusted to the needs of lexicographers. Moreover, for each lemma a direct link to the corpus examples has been provided.

After examining the output of different measures of frequency comparison, we chose to work with the difference of "frequency classes" ('Häufigkeitsklassen'; cf. Keibel 2008), a measure inspired by the word distribution in allusion to Zipf's law, which is relatively intuitive to understand and commonly used in German lexicography (cf. Klosa 2013a). The most common word in a corpus is in frequency class 0, the word(s) being approximately half as frequent as the most frequent word are in class 1, the words being approximately half as frequent as those in class 1 are in class 2, etc. We defined the shift in frequency classes of two corpora as the "difference of frequency classes" (fc_diff = fc(dereko) - fc(folk)). Since the lemma *kriegen* (en: *to get*) in FOLK is in class 5 and in DeReKo in class 12, the shift between the classes amounts to 7; hence, the difference in frequency classes (fc_diff) between the two corpora is 7 (see Table 1).

As shown in Figure 1, lemmas which have the highest fc_diff are in higher fc and are hence rare in both corpora, which is in part due to the fact that DeReKo, being much larger, contains a higher number of frequency classes (31) than FOLK (16).

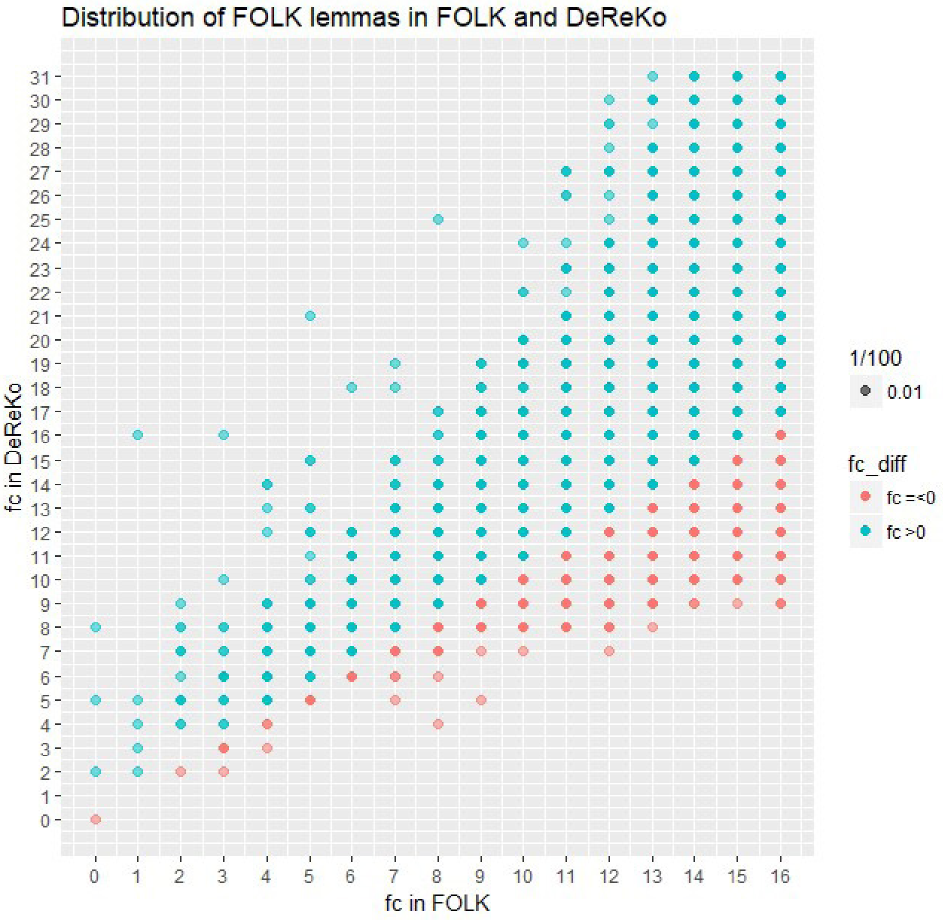


Figure 1: The distribution of the FOLK lemmas according to their frequency classes in FOLK and DeReKo. The positive fc_diff (in blue) shows the lemmas being more frequent in FOLK than in DeReKo, the negative/equal fc_diff (in red) shows the lemmas being less (or equally) frequent in FOLK than in DeReKo.

Since we wanted to concentrate particularly on lemmas being frequent in FOLK, we assumed that the relevant candidates for our headword list would be found by selecting the lemmas having the FOLK fc lower than 9, and DeReKo fc lower than 15. By setting these parameters, we were able to filter out those lemmas that have a high fc_diff but are rare in both corpora, such as, for instance, *Spinosaurus*, having a fc_diff of 7 and fc 15 in FOLK and fc 22 in DeReKo (see Table 1).

Table 1: Examples of FOLK lemmas having the fc_diff of 7 with respect to DeReKo.

Lemma	FOLK fc	DeReKo fc	fc_diff
nein	3	10	7
kriegen	5	12	7
irgendetwas	7	14	7
Schornsteinfeger	9	16	7
Transistor	11	18	7
Proseminar	13	19	7
Spinosaurus	15	22	7

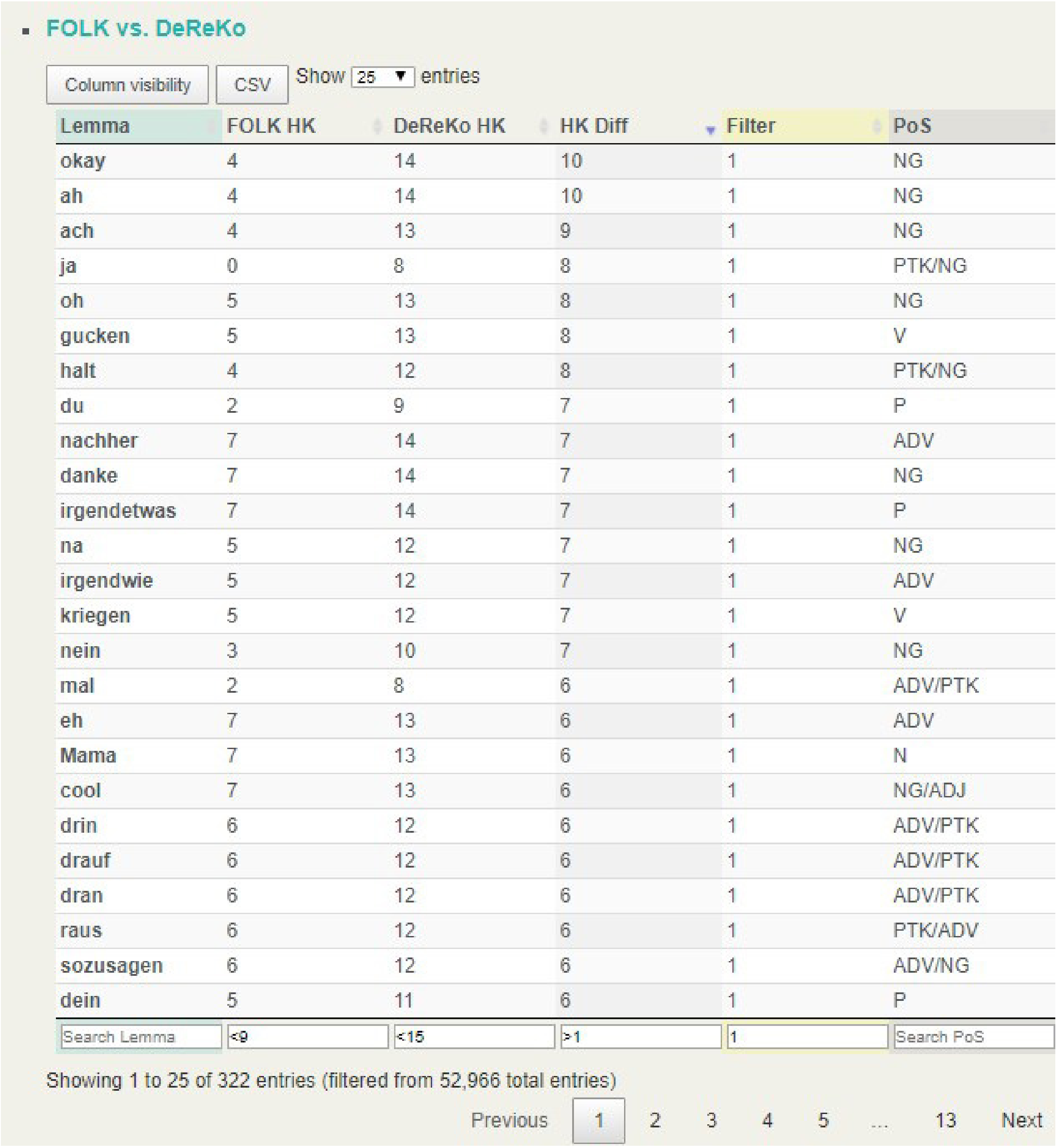
After sorting the lemma list according to the descending fc_diff, we manually examined 322 one-word lemmas whose fc_diff amounted to at least 2 (see Figure 2), with the aim to investigate whether they are suitable headword candidates for our resource. The corpus-specific lexemes that occur in very specific contexts or transcripts in FOLK, such as *Zentimeter* (maptask) and *Dialekt* (language biographical interviews), were sorted in the manual analysis.

4 Selection of headword candidates

With the method described above (see Section 3) we found lemmas that cover different subject areas of interest, such as particles, interjections, routine formulae, expressions of vagueness, deictic expressions, and other peculiarities in relation to style and register.[3] These subject areas have been mentioned before in the research literature regarding spoken German (see for example Schwitalla ^{4}2012; Fiehler 2016). The evidence for the pervasiveness of these phenomena in spoken language can be detected already by observing the top 25 lemmas with the highest fc_diff (see Figure 2). For example, the modal particles, such as *halt* and *mal* as well as interjection particles such as *ah, ach, oh*, *eh* are very common in spoken German, but insufficiently covered in German monolingual dictionaries (see Section 1). The lemmas *irgendetwas*, *irgendwie* and *sozusagen*, that also occur in the top 25 seem to testify to the prevalence of vagueness in everyday speech.

Other examples of lemmas that we identified as potential headwords among the candidates of our headword list are “passepartout” words, such as *Ding, Sache, machen, tun.* In addition, we detected several lexical alternatives to the expressions of standard written German, such as *kriegen* to *bekommen.* When inspecting the adverbs in our sample of 322 headword candidates, we found different types of deictic expressions, such as temporal (*nachher, jetzt*) and local adverbs (*da, hier*). Some adjectives like *gut*, *klar*, *cool*, *toll*, *super*, *fertig*, *geil*, *krass*, *ehrlich,* that were also detected as having high fc_diff, are also a focus of our interest because in spoken language they may also serve as discourse particles, among other functions. We have identified several verbal lemmas that can be assigned to different semantic subgroups of interest, such as perception (visual, auditory): *gucken, schauen, sehen, hören;* cognition (mental): *wissen, glauben, denken, meinen, überlegen, verstehen, kennen;* locomotion: *gehen,* communication: *reden, sagen, fragen,* emotion: *mögen, gefallen,* and

[3] A detailed overview of the subject areas can be found in Meliss & Möhrs 2017: 43.



Lemma	FOLK HK	DeReKo HK	HK Diff	Filter	PoS
okay	4	14	10	1	NG
ah	4	14	10	1	NG
ach	4	13	9	1	NG
ja	0	8	8	1	PTK/NG
oh	5	13	8	1	NG
gucken	5	13	8	1	V
halt	4	12	8	1	PTK/NG
du	2	9	7	1	P
nachher	7	14	7	1	ADV
danke	7	14	7	1	NG
irgendetwas	7	14	7	1	P
na	5	12	7	1	NG
irgendwie	5	12	7	1	ADV
kriegen	5	12	7	1	V
nein	3	10	7	1	NG
mal	2	8	6	1	ADV/PTK
eh	7	13	6	1	ADV
Mama	7	13	6	1	N
cool	7	13	6	1	NG/ADJ
drin	6	12	6	1	ADV/PTK
drauf	6	12	6	1	ADV/PTK
dran	6	12	6	1	ADV/PTK
raus	6	12	6	1	PTK/ADV
sozusagen	6	12	6	1	ADV/NG
dein	5	11	6	1	P
Search Lemma	<9	<15	>1	1	Search PoS

Showing 1 to 25 of 322 entries (filtered from 52,966 total entries)

Figure 2: Top 25 lemmas with the highest fc_diff between FOLK and DeReKo, having FOLK fc (*FOLK HK*) lower than 9, DeReKo fc (*DeReKo HK*) lower than 15, and fc_diff of at least 2 classes. *HK Diff* stands for fc_diff; *Filter:1* stands for the lemmas we consider for the headword list (no numerals, no proper names, etc.); *PoS* stands for the aggregation of parts-of-speech with which a lemma occurs in FOLK.

modal verbs: *können, müssen*. Some qualitative studies demonstrate that the verbs of these semantic subgroups have an important potential in interactional contexts to realize several special meanings, functions and formal particularities according to the different subject areas described above (cf. Deppermann et al. 2017). Additionally, they can be the lexical basis of several multi-word-lemmas, which must be considered in further work.

After defining the headword candidates based on the corpora comparison, we analyze the corpus examples and focus on the peculiarities in meaning, use and function in talk-in-interaction by following the approaches of interactional linguistics and lexicology.

5 Discussion

The selection of headwords is related to the scope of the dictionary and its static vs. dynamic "dictionary under construction" conception (on these aspects, cf. Atkins & Rundell 2008: 160 ff.; Klosa 2013: 518; Schnörch 2005: 71; Wiegand 1983). Since our lexicographic resource is intended to be a digital one, in terms of quantity, our list of headwords is per definition dynamic and extensible. As shown in Section 4, for the first step of dictionary creation we used the most common one-word lemmas with a prominent frequency difference between FOLK and DeReKo. The combination of automated procedures and manual analysis of the data has proven to be an effective and sustainable way of approaching the task of headword selection. With the help of the tool for automatic sorting and filtering of the headword candidates according to their parts of speech, frequency and other features, the selection of more headwords can be carried out in a transparent and scalable way in further steps of our project, and can be enhanced in significantly.

The measure of frequency class difference proved to be a suitable measure for detecting the difference in the lemma distribution in the two corpora. Since we integrated the other measures that were commonly used for lemma frequency comparison as well, we can further investigate the effects of other measures and compare them with each other in further studies.

We did not encounter any great discrepancies that emerged from different lemmatization conventions in the two corpora, presumably because we worked with frequent lemmas and closely comparable operationalizations. However, the lemmatization process and the varieties of conventions in the German language corpora are certainly an issue that has to be taken into consideration when comparing the sets of lemmas in two corpora in future works.

The detection of one-word lemmas is only the first step towards a set of headwords needed to represent the lexicon of spoken German in interaction, which is also built upon multi-word expressions to a significant degree. In further steps of our project, we will work on the lexicographic implementation and the detection and integration of such expressions into our headword concept.

References

Atkins, B. T. S., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Bergmann, P. (2017). Gebrauchsprofile von *weiß nich* und *keine Ahnung* im Gespräch - Ein Blick auf nicht-responsive Vorkommen. In H. Blühdorn, A. Deppermann, H. Helmer, T. Spranz-Fogasy (eds.) *Diskusmarker im Deutschen. Reflexionen und Analysen*. Göttingen: Verlag für Gesprächsforschung, pp. 157-182.

Deppermann, A., Proske, N., Zeschel, A. (eds.) (2017). *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch*. Tübingen: Narr (= Studien zur deutschen Sprache, Band 74).

Eichinger, L. M. (2017). Gesprochene Alltagssprache. In *Deutsche Akademie für Sprache und Dichtung / Union der deutschen Akademien der Wissenschaften* (eds.) Vielfalt und Einheit der deutschen Sprache. Zweiter Bericht zur Lage der deutschen Sprache. Tübingen: Stauffenburg, pp. 283-331.

Fiehler, R. (2016). Gesprochene Sprache. In: A. Wöllstein (ed.) (2016): *Duden – Die Grammatik. Unentbehrlich für richtiges Deutsch*. Berlin: Dudenverlag, pp. 1181-1260.

FOLK. *Forschungs- und Lehrkorpus Gesprochenes Deutsch* (Release 2.8 vom 06.04.2017). Accessed at: http://agd.ids-mannheim.de/folk.shtml. [16/05/2018].

Hansen, C., Hansen, M. H. (2012). A Dictionary of Spoken Danish. In R. V. Fjeld & J. M. Torjusen (eds.) *Proceedings of the 15th EURALEX International Congress. 7-11 August 2012.* Oslo, Norway: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 929-935.

Institut für Deutsche Sprache (2017). *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2017-I* (Release 08.03.2017). Mannheim: Institut für Deutsche Sprache. Accessed at: http://www.ids-mannheim.de/direktion/kl/projekte/korpora/releases.html. [16/05/2018].

Keibel, H. (2008). *Mathematische Häufigkeitsmaße in der Korpuslinguistik: Eigenschaften und Verwendung*. Mannheim: Institut für Deutsche Sprache. Elektronische Ressource.

Klosa, A. (2013a). Aktuelle Tendenzen in der deutschen Lexikographie der Gegenwart. In G. Stickel, T. Váradi (eds.) *Lexical Challenges in a Multilingual Europe. Contributions to the Annual Conference 2012 of EFNIL in Budapest*. Frankfurt am Main: Lang, pp. 75-93. (= Duisburger Arbeiten zur Sprach- und Kulturwissenschaft 99).

Klosa, A. (2013b). The lexicographical process (with special focus on online dictionaries). Berlin/New York. de Gruyter. In R. H. Gouws, U. Heid, W. Schweickard, H. E. Wiegand (eds.) *Wörterbücher. Ein internationales Handbuch zur Lexikographie*, pp. 517-524 (= Handbücher zur Sprach- und Kommunikationswissenschaft. Bd. 5.4).

Kupietz, M., Keibel, H. (2009). The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In M. Minegishi, Y. Kawaguchi (eds.) *Working Papers in Corpus-based Linguistics and Language Education, no. 3*. Tokyo: Tokyo University of Foreign Studies (TUFS), pp. 53-59. Accessed at: http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf. [16/05/2018].

Meliss, M. (2016). Gesprochene Sprache in DaF-Lernerwörterbüchern. In B. Handwerker, R. Bäuerle, B. Sieberg (eds.) *Gesprochene Fremdsprache Deutsch*. Baltmannsweiler: Schneider, pp. 179-199. (= Perspektiven Deutsch als Fremdsprache, Band 32).

Meliss, M., Möhrs C. (2017). Die Entwicklung einer lexikografischen Ressource im Rahmen des Projektes LeGeDe. In *Sprachreport* 4/2017, pp. 42-52. Accessed at: http://pub.ids-mannheim.de/laufend/sprachreport/pdf/sr17-4.pdf. [16/05/2018].

Meliss, M., Möhrs, C., Ribeiro Silveira, M. (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*. 68/1, pp.103-138.

Möhrs, C., Meliss, M., Batinić, D. (2017). LeGeDe - towards a corpus-based lexical resource of spoken German. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, V. Baisa (eds.) *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, pp. 281-298.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK. Accessed at: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf. [16/05/2018].

Schmidt, T. (2014). The research and teaching corpus of spoken German – FOLK. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds.) *Proceedings of the 9th Conference on International Language Resources and Evaluation (LREC'14)*. 26-31 May 2014. Reykjavik, Iceland, pp. 383-387. Iceland: ELRA. Accessed at: http://www.lrec-conf.org/proceedings/lrec2014/index.html. [16/05/2018].

Schnörch, U. (2005). Die elexiko-Stichwortliste. In U. Haß (ed.) *Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz*. Berlin/New York: de Gruyter, pp. 71-90. (= Schriften des Instituts für Deutsche Sprache 12).

Schwitalla, J. (^{4}2012): *Gesprochenes Deutsch. Eine Einführung*. Berlin: Schmidt. (= Grundlagen der Germanistik, Band 33).

Siepmann, D. (2015). Dictionaries and spoken language: A corpus-based review of French dictionaries. In *International Journal of Lexicography*. 28 (2), pp. 139-168.

Stadler, H. (2014). Die Erstellung der Basislemmaliste der neuhochdeutschen Standardsprache aus mehrfach linguistisch annotierten Korpora. (= *OPAL - Online publizierte Arbeiten zur Linguistik* 5/2014). Mannheim: Institut für Deutsche Sprache. Accessed at: http://pub.ids-mannheim.de/laufend/opal/opal14-5.html. [16/05/2018].

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

Verdonik, D., Sepesy Maučec, M. (2017). A Speech Corpus as a Source of Lexical Information. In *International Journal of Lexicography*. 30 (2), pp. 143-166.

Westpfahl, S. (2014). STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In L. Levin, M. Stede (eds.) *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, pp. 1-10.

Westpfahl, S., Schmidt, T. (2016). FOLK-Gold – A GOLD standard for Part-of-Speech-Tagging of Spoken German. In N. Calzolari (Conference Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds.) *Proceedings of the Tenth Conference on International Language Resources and Evaluation (LREC'16),* Portorož, Slovenia. Paris: European Language Resources Association (ELRA), pp. 1493-1499.

Wiegand, H.-E. (1983). Was ist eigentlich ein Lemma? Ein Beitrag zur Theorie der lexikographischen Sprachbeschreibung. In H.-E. Wiegand (ed.) *Studien zur neuhochdeutschen Lexikographie III*. Hildesheim et al.: Olms, pp. 401-474. (= Germanistische Linguistik, 1-4/82).