# A Corpus-Based Lexical Resource of Spoken German in Interaction

## Meike Meliss[1], Christine Möhrs[2],

## Maria Ribeiro Silveira[2], Thomas Schmidt[2]

[1] Leibniz-Institut für Deutsche Sprache, Mannheim /
Universität Santiago de Compostela (Spanien)
[2] Leibniz-Institut für Deutsche Sprache, Mannheim
E-mail: meliss@ids-mannheim.de / meike.meliss@usc.es, moehrs@ids-mannheim.de,
silveira@ids-mannheim.de, thomas.schmidt@ids-mannheim.de

## Abstract

This paper presents the prototype of a lexicographic resource for spoken German in interaction, which was conceived within the framework of the LeGeDe-project (LeGeDe=Lexik des gesprochenen Deutsch). First of all, it summarizes the theoretical and methodological approaches that were used for the initial planning of the resource. The headword candidates were selected by analyzing corpus-based data. Therefore, the data of two corpora (written and spoken German) were compared with quantitative methods. The information that was gathered on the selected headword candidates can be assigned to two different sections: *meanings* and *functions in interaction.*
Additionally, two studies on the expectations of future users towards the resource were carried out. The results of these two studies were also taken into account in the development of the prototype. Focusing on the presentation of the resource's content, the paper shows both the different lexicographical information in selected dictionary entries, and the information offered by the provided hyperlinks and external texts. As a conclusion, it summarizes the most important innovative aspects that were specifically developed for the implementation of such a resource.

**Keywords:** online lexicography; spoken German; corpus-based

## 1. Introduction

The lexicographic resource described in this article in its conception and implementation was conceived and created in the research project "Lexik des gesprochenen Deutsch" (=LeGeDe) between 2016 and 2019 at the Leibniz Institute for the German Language (IDS) in Mannheim[1]. The cooperation between the Department of Pragmatics and the Department of Lexical Studies at the IDS enabled a connection

---

[1] The resource is created within the framework of the third-party funded research project LeGeDe financed by the Leibniz Association (Leibniz Competition 2016, Funding line: 1: Innovative projects). Project website: http://www1.ids-mannheim.de/lexik/lexik-des-gesprochenen-deutsch.html.

of the corresponding professional competence necessary for the creation of a corpus-based lexicographic resource of spoken German in interaction during the project period. The creation of such a corpus-based electronic resource of spoken German, based on the one hand on research on the peculiarities of spoken vs. written language use, and on the other hand on important experience in the field of electronic lexicography (cf. Klosa & Müller-Spitzer, 2016), was the project's main objective. Both from the research's point of view on spoken language and from a lexicographical perspective, a completely new form of lexicographic language description and presentation needed to be developed. Furthermore, it was necessary to generate novel lexicographic types of information with audio-features that refer to the function of lexical units in interactional contexts, for which so far hardly any lexicographical models exist. The lexicographical prototype is intended to primarily serve as a knowledge repository and vocabulary documentation (https://www.owid.de/legede/). The resource addresses scientists, interactional linguists, and lexicologists as its primary target group (cf. Meliss et al., 2018b, 2019). Nevertheless, we are convinced that learners of German can also benefit from the resource if the experts take the corresponding intermediate position. For this purpose, quantitative and qualitative methods were developed with which the specifics of the spoken-language lexicon of German could be identified, analysed, and prepared for lexicographical application on the basis of oral corpora created at the IDS (cf. the program area "Oral corpora").

In this paper we present the most important challenges and results of the LeGeDe-project. Therefore we introduce in section 2 the project's background (research questions, aims, and objectives), and in section 3 we show the relevant information about our corpus-based database. In section 4 we present some relevant results of two empirical studies on expectations we carried out at the beginning of the project. The information on lexicographical implementation is presented in section 5, using illustrative examples. In our concluding remarks (cf. section 6), we emphasize the innovative aspect of the LeGeDe-resource and give a brief outlook on further research and work areas.

## 2. Research questions and objectives

The LeGeDe-project is based on the following four main assumptions and observations:

(i)     There are differences at several linguistic levels between spoken and written German. With regard to the lexicon, the divergences can have an effect on both the lexical inventory and the relation with its form, meaning, and use (cf. Deppermann et al., 2017; Fiehler, 2016; Imo, 2007; Schwitalla, 2012).

(ii)    The way existing dictionaries codify the characteristics of the spoken German lexicon is deficient in several ways (cf. e.g. Meliss, 2016; Meliss et al., 2019; Moon, 1998; Trap-Jensen, 2004). There are currently hardly any corpus-based lexicographic projects that aim to develop a lexicon of spoken language. Only one small project on interjections (cf. Hansen & Hansen, 2012) was

carried out on Danish. The results of two LeGeDe-surveys on the expectations and requirements of a lexicographic resource for the specifics of spoken German (cf. Meliss et al., 2018b, 2019), carried out in cooperation with the project "Empirische Methoden", confirm that the lexicographical codification of spoken language and its interactional features are not satisfactorily taken into account in the currently existing dictionaries (cf. Meliss, 2016: 195; Eichinger, 2017: 283). Despite some recent advances in corpus-based lexicography of spoken language (cf. Verdonik & Sepesy Maučec, 2017; Hansen & Hansen, 2012; Siepmann, 2015), experience with spoken language data in lexicography has so far been rather rare. Therefore, the LeGeDe-resource can hardly rely on existing models that could serve as guidance for the compilation of a suitable list of headwords and for the lexicographical modelling and implementation.

(iii)    The need for information on typical spoken vocabulary has increased in general and in various areas of application, e.g. in learning and teaching areas (especially in secondary education and in the areas of German as a foreign and/or second language) as well as in the research and publication area in connection with the production of suitable study materials (cf. Handwerker et al., 2016; Imo & Moraldo, 2015; Meliss & Möhrs, 2018; Moraldo & Missaglia, 2013; Reeg et al., 2012; Sieberg, 2013). For example, in the "Common European Framework of Reference for Languages" (=GeR), among other items on the assessment grid for oral communication and the parameter "interaction" for level C1, it was explicitly noted that a learner should be able to choose an appropriate turn from a repertoire of means of discourse in order to make his utterance appropriate (cf. Trim et al., 2001: 37).

(iv)    In addition, the results of the empirical studies, carried out in the LeGeDe-project show that more than 70% of L1 and L2 speakers of German expressed a need for a dictionary on specifics of spoken German. This observation confirms the basic assumption of an increasing demand for such a resource.

These basic assumptions are the starting points for conceptual considerations in order to develop our lexicographical resource and lead to the following essential theoretical, methodological, and application-oriented aspects, which arose when dealing with the topic in the project work:

- development of quantitative and qualitative methods to identify spoken-language lexical elements and their specific characteristics in interactional contexts in comparison to the lexicon of written language (cf. Meliss & Möhrs, 2017),

- preparation of a list of headword candidates and selection of suitable lemmas for the prototype of the LeGeDe-resource (cf. Meliss et al., 2018a),

- development of further (corpus-)linguistic methods for analysing and structuring

spoken language data, also for structuring automatically generated corpus-based data (cf. Möhrs et al., 2017),

- determination of the peculiarities of spoken language usage at different levels (form, content/function, conversational setting etc.), in our project with a focus on lexical specifics,

- development of innovative forms of lexicographical information, which refer to the function of lexical units in interactional contexts (taking into account transcripts and their associated audios).

## 3. Database of the LeGeDe-project

The studies on the research object of the LeGeDe-project are carried out exclusively on the basis of the "Research and Teaching Corpus of Spoken German" (=FOLK: cf. Schmidt, 2014a; Kupietz & Schmidt, 2015). FOLK is the largest corpus of conversational German, which was developed at the IDS and is integrated in the "Database for spoken German" (=DGD: cf. Schmidt, 2014b). FOLK primarily contains authentic data from interactive conversations (cf. Schmidt, 2017). Included are conversation recordings and transcripts (partly also video recordings) from German-speaking regions in various private, institutional, and public contexts. The data can be categorized by the following characteristics: oral media, authentic, spontaneous, mostly of the standard language, and up-to-date. Currently, FOLK is available in DGD version 2.12 with almost 250h/2.4 million tokens and 306 different speech events.[2] As a corpus analysis tool, the DGD offers a variety of possibilities for indexing oral data according to linguistic and interactional characteristics, and is constantly further developed and equipped with innovative corpus technology functionalities. Structured token searches can be realized via the user interface and searched via four annotation levels (cGAT transcript, normalization, lemmatization, PoS). In addition, metadata on speakers and on the conversation event can be retrieved for the conversations. The size of the corpus, the data it contains from authentic interaction, and the annotation of the data provide a reliable basis for lexicological and interactional analysis.

Since 2018, the use of the tool *Lexical Explorer* (cf. Batinić-Lemmenmeier, in press), an application developed during the LeGeDe-project, allows further access to FOLK as well as to the GeWiss ("Gesprochene Wissenschaftssprache") corpus. With this tool, quantitative corpus data on spoken German can be explored with the help of frequency tables regarding the distribution across word form variation, co-occurrences, and metadata.

---

[2] The samples analyzed in the LeGeDe-project were based on DGD Version 2.11.

## 4. Empirical studies: expectations on a dictionary

## of spoken German

Two empirical studies were carried out at the beginning of the LeGeDe-project. The main goal of these studies was to shed light on people's expectations on the planned lexicographical online-resource. In the first study, selected experts were polled in the form of a guided interview. In the second, a broader online survey was conducted, which aimed to reach a wider range of potential users.[3] With our two conducted surveys (interview and online survey) we intended to learn about expectations with regard to as many different lexicographical aspects as possible. In addition, sociodemographic data were also collected, and questions concerning the personal handling and use of (online) dictionaries together with the specific handling of the spoken-language lexicon were asked.

In our first study, we interviewed 17 experts from different linguistic areas. Each interview consisted of 30 questions mainly in an open question format, so the analysis of the greater part of the data was performed with qualitative methods. A smaller number of questions were presented in a closed format, so these data could be analysed with quantitative methods and be compared to answers from the online survey. Nevertheless, when viewing the results of the interview and especially when comparing them to data from the online survey, it must be considered that these are data from only 17 participants. The purpose of the online survey was also to ask for the opinions of a wider range of potential users and beneficiaries (e.g. linguists, teachers of German, domestic or abroad) of the planned resource. For this questionnaire, which contained 35 questions, we mainly used closed question formats. Altogether 333 participants completed the online survey.

In the following sections we present particular results relevant for basic considerations for the implementation of the LeGeDe-resource as well as results directly concerning it.

### 4.1 Target group of the planned resource

The question of the target group is fundamental for the lexicographical implementation of the collected data. Since the LeGeDe-resource initially functions as a knowledge store and vocabulary documentation, the presentation of the data is primarily geared towards a scientifically interested group of users (including conversation researchers, interaction linguists, corpus linguists, lexicologists, lexicographers).

However, the results of our empirical surveys on the question *For which target group*

---

[3] In Meliss et al. (2018b, 2019), the results from both studies are summarized either from a general or from a L1 vs. L2 perspective.

*could a dictionary of spoken German be of particular interest?*[4] have also shown that users in certain learning situations – especially in speech production situations – could benefit from the LeGeDe-resource. For this purpose it would be necessary that the experts (scientists, teachers, etc.) take on a corresponding mediating position. Based on the data provided by the LeGeDe-resource, language teaching material for the concrete treatment of specific lexical phenomena in spoken interaction could be developed from an application-oriented perspective for German as a foreign or mother tongue language (cf. Meliss et al., 2018b: 132; 2019: 116).

## 4.2 Headword candidates

We first look at the results for the following question of the online survey: *What kind of headwords would you expect in a dictionary of spoken German?* Different observations result from the answers given by the test subjects: (i) Most of the online survey respondents (87.8%) expect headwords which have a different meaning and functionality in spoken interaction than in written use. (ii) In a dictionary of spoken German, respondents to the online survey expect headwords to have a formulaic use (79.5%) as well as headwords with a special combination potential (e.g. patterns, specific units, etc.; 74.6%). (iii) Headwords that are exclusively spoken (77.7%) and those that occur particularly frequently in spoken interaction (71.6%) are also desired by the test subjects of the online survey. (iv) Slightly more than half of the online survey participants also expect lexical units that can be characterized by formal phonetic contraction (57.5%). A look at the respondents' responses to "Miscellaneous" shows that, among other things, headwords with a different spectrum of linguistic variation are also desired.

According to the experts' assessment, lexical units with a different combination potential in spoken vs. written language are the most desired headwords (94.1%). In their opinion, this includes constructions, lexical expressions, syntagmatic combinations, formulas, etc., as well as multiword lemmas. The experts also listed lexical units with differences in meaning or function as important headwords.

## 4.3 Information on the headword candidates

This section looks at the answers of our online survey to the following question: *In your opinion, what information should be offered in a dictionary of spoken German?* From the five different answer possibilities to this question: *Definitely* (1), *Useful, but not absolutely necessary* (2), *Not useful, but nevertheless desirable* (3), *Unnecessary* (4) to *I don't know* (5), options 1-4 are shown in Fig. 1.

---

[4] The interview and the survey were conducted in German. Questions and answers are translated into English for better comprehension.

The results of the online survey – visualized here by the median and the arithmetic mean (=AM) – show that a broad spectrum of information, namely on pronunciation, meaning/function in context, formal peculiarities and features in combinatorics and word formation, together with the range of corpus data, metadata on the conversation situation, and comparative information (written vs. spoken) was equally evaluated by the respondents with the answers *Definitely* or *Useful, but not absolutely necessary*. It is also notable that the participants of the online survey on the topic of information provision also asked for information on frequency and style, index, and diatopic distribution. The evaluation of the answers must be considered in conjunction with those from the question about possible headwords (cf. section 4.2).
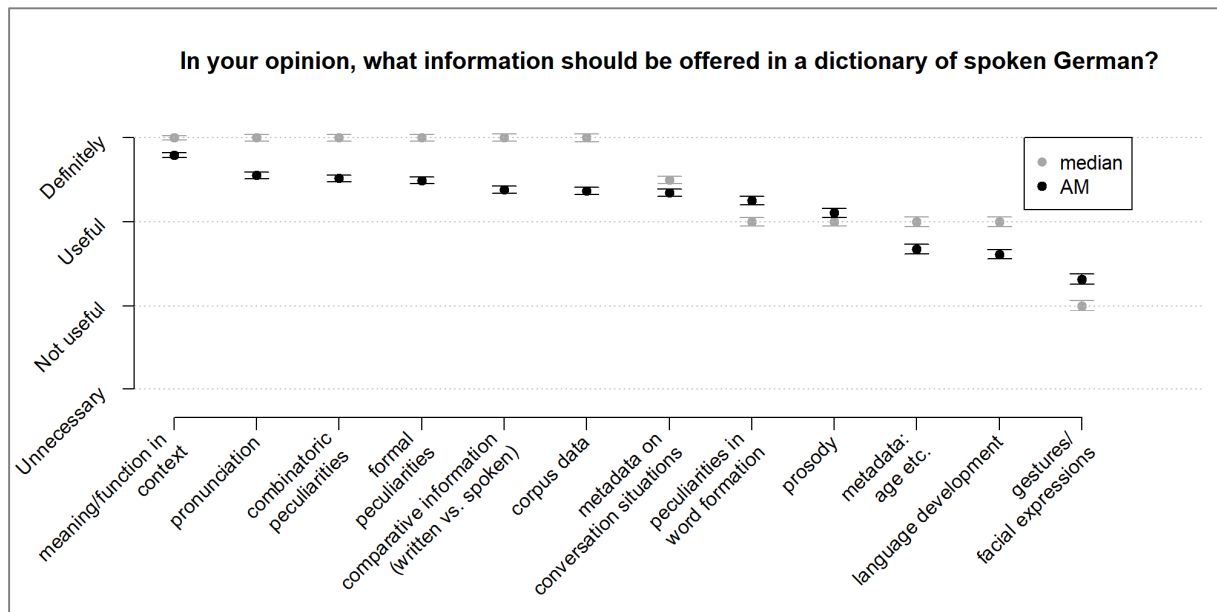


Figure 1: Distribution of expectations regarding the information provided (online survey).

When using the results of the expert interviews for comparison, it becomes clear that the information on pronunciation, meaning in context, special features in form, special features in combinatorics, supply of corpus documents, metadata on the conversation situation, and prosody were rated equally highly as *Definitely* (cf. Fig. 2). In addition, the experts – similar to the respondents to the online survey – also mentioned information on linguistic variation.

A comparison of the two surveys allows the following conclusions to be drawn: There are similarities in the following points: (i) Most of the information is rated by all respondents as necessary and useful without major differences. (ii) An exception is information on metadata, such as age, language development, and gestures/facial expressions, which have been classified as *Not useful, but nevertheless desirable*. Differences between the two surveys lie mainly in the information provided on prosody (for the experts *Definitely*, for the respondents of the online survey *Useful, but not absolutely necessary*). This divergence can be explained by the higher degree of specific conversational linguistic expertise of the interviewees from the expert interviews (cf.

Meliss et al. 2018b: 126-128, 2019: 104-106).
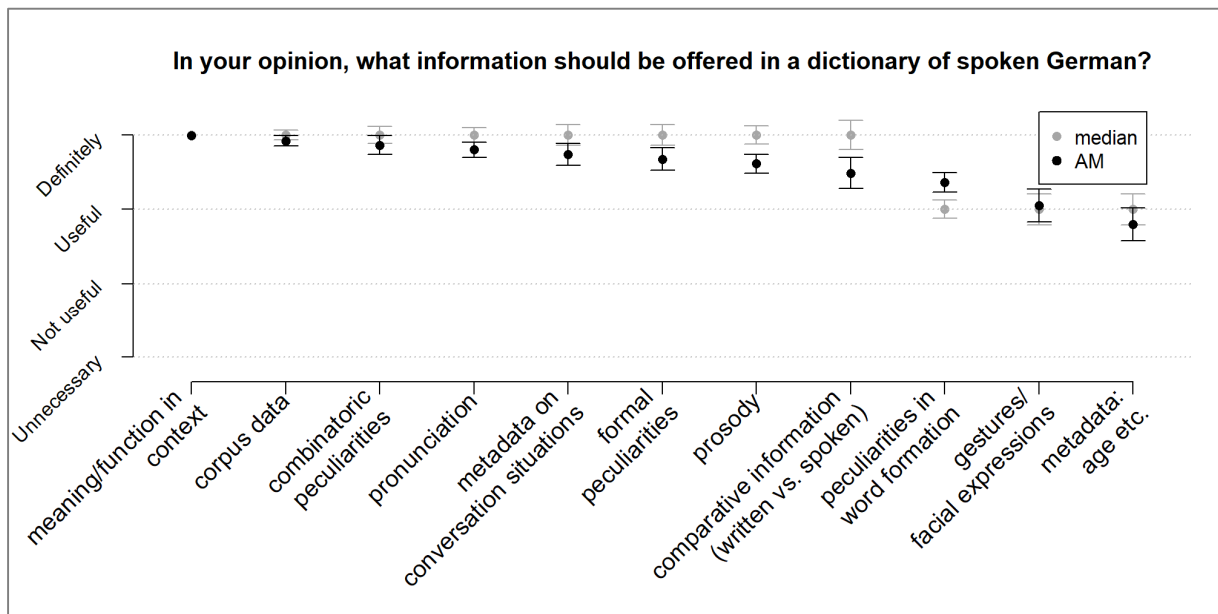


Figure 2: Distribution of expectations regarding the information provided (expert interviews).

In the following, the information included in the prototype of the resource is explained in more detail. Many of the expected aspects could be taken into account in the lexicographical implementation.

## 5. The LeGeDe-resource

The LeGeDe-resource offers an extensive range of information for each headword. The result is a complex lexicographical structure. In the following sections we explain the design and implementation of five aspects: (i) the identification of headword candidates, and the lemmas described in the dictionary (cf. 5.1), (ii) the range of information for each headword (cf. 5.2), (iii) the outer texts (cf. 5.3), (iv) the linking of the dictionary articles with the DGD, and (v) the possibility of further corpus analysis (cf. 5.4).

### 5.1 Headword candidates

One of the key research and methodological issues that the LeGeDe-project has addressed is related to the identification of typically spoken lexical peculiarities, and thus to the comparison with dictionaries based on written language. In direct relation to the distinctive features of lexical peculiarities on written and spoken language in interaction, a list of headword candidates for the LeGeDe-resource is drawn up (cf. Meliss et al., 2018a). As typical phenomena of spoken language, these candidates are used in spontaneous interaction and thus are clearly distinguishable from written language aspects.

The considerations regarding one-word lemmas, which have a specific meaning and function in interaction (e.g. interjections), have to be complemented with the integration of multiword expressions and constructions with specific functions in interaction (e.g. *was weiß ich* [engl. *I don't know*], *keine Ahnung* [engl. *no idea*], *guck mal* [engl. *look!*]) as headword candidates (cf. e.g. Bergmann, 2017; Günthner, 2017, Helmer & Deppermann, 2017; Helmer et al., 2017; Imo, 2007; Zeschel, 2017).

Hence, a corpus-based and interpretative method was developed in the LeGeDe-project (cf. Meliss et al., 2018a) to create a list of headwords, with which the most important candidates of the typical spoken lexicon could be uncovered in interaction (cf. with regard to the expectations on the headword candidates, the results are shown in section 4.2). For the comparison with the written language, the German reference corpus (=DEREKO, version 2017 I, cf. Kupietz & Keibel, 2009; Kupietz et al., 2018) was used. The method applied is briefly explained below.[5]

Since we wanted to use DEREKO as a representation of current written language, we have excluded data that contain the conceptually spoken language presented in Wikipedia discussions as well as the subcorpus "Sprachliche Umbrüche" from the years 1945 to 1968. One of the steps was to calculate the difference in lemma distribution in the two corpora by using different effect measures (odds ratio, %diff, relative risk, binary protocol of relative risk and frequency classes) and measures of statistical significance (log likelihood ratio and chi square). The lemma comparison table has been integrated into a tool we developed to quickly and easily filter and sort the data. With the help of this tool, the headword candidates can be dynamically evaluated, executed, and explored, and the parameters can be adapted to the needs of the lexicographers. After examining the results of different measurements of the frequency comparison, we opted for the difference of the "frequency classes" ("Häufigkeitsklasse" = HK; cf. Keibel, 2008, 2009), a measurement which is relatively intuitive to understand and frequently used in German lexicography (cf. e.g. Klosa, 2013). The most common word in a corpus is in frequency class 0, whereas the word(s) in class 1 is (are) about half as common as the most common word(s) in class 0, the words in class 2 are about half as common as those in class 1, etc. We calculated the difference of the frequency classes of a lemma in the two corpora as "difference of the frequency classes" (fc_diff = fc(dereko) – fc(folk)). After sorting the lemma list by descending fc_diff, we extracted about 320 one-word lemmas whose fc_diff was at least 2. The manual check of these candidates enabled us to see if they were suitable headword candidates in the one-word lemma range for our resource. Table 1 shows the top 25 candidates for which we can define different headword groups.

---

[5] For details cf. Meliss et al. (2018a).

| No. | Lemma | FOLK HK | DeReKo HK | HK Diff |
|---|---|---|---|---|
| 1 | ah | 4 | 14 | 10 |
| 2 | okay | 4 | 14 | 10 |
| 3 | ach | 4 | 13 | 9 |
| 4 | ja | 0 | 8 | 8 |
| 5 | irgendetwas | 6 | 14 | 8 |
| 6 | gucken | 5 | 13 | 8 |
| 7 | oh | 5 | 13 | 8 |
| 8 | halt | 4 | 12 | 8 |
| 9 | irgendwie | 4 | 12 | 8 |
| 10 | du | 2 | 9 | 7 |
| 11 | danke | 7 | 14 | 7 |
| 12 | nachher | 7 | 14 | 7 |
| 13 | kriegen | 5 | 12 | 7 |
| 14 | na | 5 | 12 | 7 |
| 15 | nein | 2 | 10 | 7 |
| 16 | also | 2 | 8 | 6 |
| 17 | Hey | 8 | 14 | 6 |
| 18 | runter | 7 | 13 | 6 |
| 19 | wieso | 7 | 13 | 6 |
| 20 | cool | 7 | 13 | 6 |
| 21 | Ahnung | 7 | 13 | 6 |
| 22 | Mama | 7 | 13 | 6 |
| 23 | drin | 6 | 12 | 6 |
| 24 | sozusagen | 6 | 12 | 6 |
| 25 | dein | 5 | 11 | 6 |

Table 1: TOP 25 of one-word lemmas from a statistical point of view (FOLK, Release 2.11, cf. *Lexical Explorer*: "Study corpus vs. DeReKo", Study corpus HK = <9, DeReKo HK = <15, HK Diff = >1, Filter = 1).

Headword candidates as one-word lemmas are defined on the basis of this method. Manual analysis is used to record information on very different grammatical, semantic, and interactional linguistic aspects. For each one-word lemma, a sample of 300 hits is drawn from FOLK. Of these, 100 valid (i.e. clear audio) hits are analysed and coded in detail. The range of information on the headwords is explained in more detail in section 5.2.

The further step to analyse the sample of each selected headword according to formal, semantic, syntactic and functional criteria shows, among other things, whether there are any occurrences of the lemma in the data that refer to one of the meanings of the

one-word lemma (e.g. 'abwarten' [engl. *to wait*] as one of the basic meanings of the lemma *gucken* [engl. *to look*], selected as one of our headwords (cf. no. 6 in Table 1). The results on the meaning-based analysis of one-word lemmas lead to a dictionary article "Bedeutungen" [engl. "Meanings"] (= module 1), which we describe in more detail in section 5.2.2.

In addition, the detailed analysis work on the sample shows the possibility of the occurrence of units with a special interactional function. These can be one-word or multi-word units related to the list of identified headword candidates (e.g. *halt* as a 'modal particle' cf. no. 8, or *guck mal* as a 'discourse marker' cf. no. 6 in Table 1). Section 5.2.3 describes the lemmas with interactional functions in more detail.

## 5.2 Range of information on the headwords

In the following, the central lexicographic information sections (overview, module 1, module 2) on selected headwords, which have been edited accordingly, will be presented (with regard to the expectations on the headword candidates, cf. the results of the studies in section 4.3).[6]

### 5.2.1 General overview

For each headword, general overview information is available and offers, in a descriptive form, meaning- and function-oriented information (e.g. *eben* [engl. *just*[7]], cf. Fig. 3).

A clear modular division of the information enables the presentation of lexical-semantic information on the one hand, which is oriented to the respective meaning of the corresponding senses of a lemma (= module 1), and on the other hand of function-specific interactionally oriented information (= module 2). For both areas, specific lexicographical information was used or newly developed for description purposes, which offers completely new insights and formats in addition to traditional dictionary information.

Different cross-connections between the two modules are made explicit by an internal link. An extended external information offer is provided by a link that leads to further lexicographic resources (e.g. DWDS) on the one hand and to FOLK and the *Lexical Explorer* on the other hand. In addition, the calculated corpus-based frequency class difference between the headwords in the respective corpora (written: DEREKO, spoken: FOLK) is visualized (cf. Fig. 4).

---

[6] This figure and also the following (with excerpts from the resource) are based on the beta version of the resource (last update: 5 August 2019).
[7] *eben* as an adverb can generally be translated as *just* in English. For *eben* as modal or discourse particles, there are contexts in English in which *just* could also be used. But a clear lexical equivalent of the particles *eben* does not exist in English.
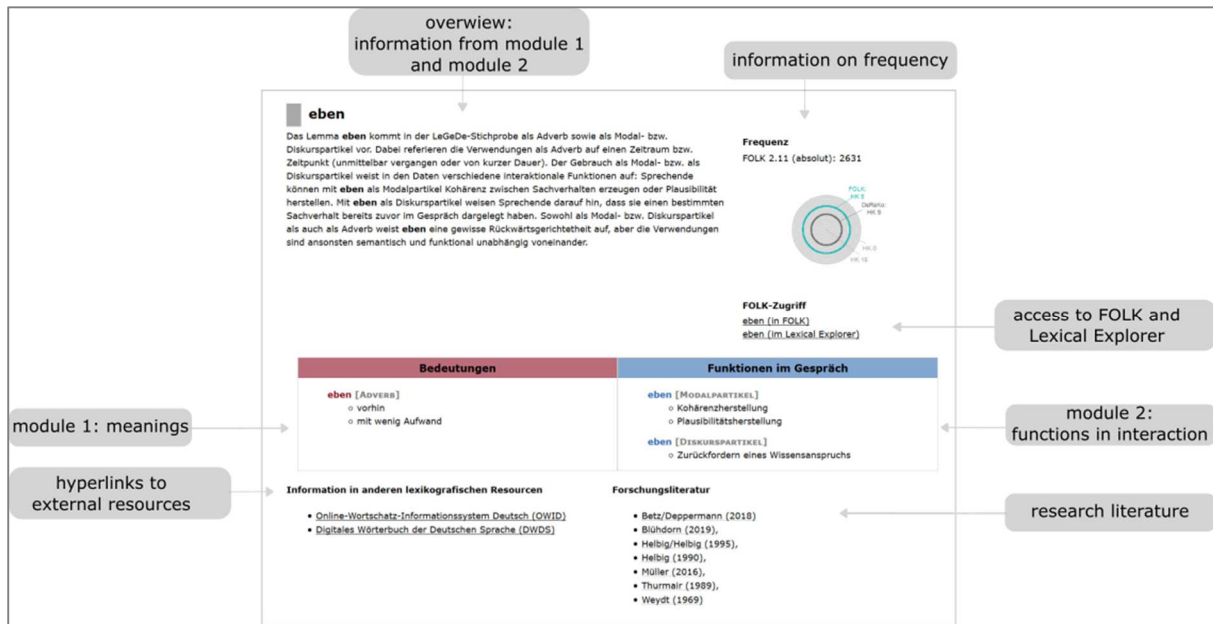
Figure 3: Overview article of the lemma *eben* (screenshot).
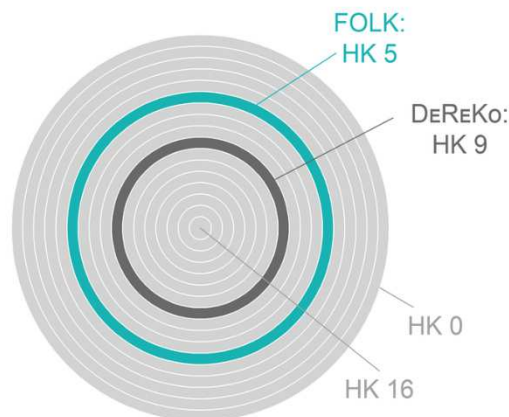


Figure 4: Visualization of the frequency classes of *eben* in FOLK and DeReKo.

An optional short reference to the related research literature enables an insight into relevant sources for each headword.

### 5.2.2 Module 1: Meanings

In addition to relevant general **sense-independent information (1)** (e.g. word class: adverb, verb, noun; morpheme structure in case of lexical compounds or affix constructions: *Ahnung*: Basis: *ahn-* (=verbal stem), *-ung* (=derivative suffix) [engl. *idea/knowledge*]); formal variation (e.g. *gucken*: <kucken>/[kucken]); research literature) the data of module 1 is mainly supplemented by information on meaning and combinatorics. The different information items of module 1 (see **(1)-(9)**) are

subsequently explained using the verb *wissen* [engl. *to know*] as an example and presented at a glance in Fig. 5.

(i) The **sense-related information** of each lemma is semantically identified by a **short label (2)** for disambiguation of meaning and by short **semantic paraphrases (3)**. In addition, a **transcript box (4)** with a transcript title, a short description of the context, an illustrative transcript excerpt (+audio), and an optional commentary is offered (cf. Fig. 5).

(ii) **Formal peculiarities**: There is also the possibility of pointing out formal peculiarities in a short comment. Different aspects can be commented on, such as the distinctive combinatorial behaviour with modal verbs, the use of certain verbal modes or the connection with certain particles or deictic expressions.

(iii) **Combinatorics (5)**: In conjunction with information on combinatorics, we distinguish between **structural patterns (6)**, **fixed phrases/collocations (7)**, and **interactional units (8)**. Transcript boxes illustrate the corresponding phenomena (cf. Fig. 5).

a. The **structural patterns (6)** (= Strukturmuster) are offered in an abstract, formulaic way (e.g. <jemand weiß, dass/ob etwas der Fall ist//was der Fall ist> [engl. <someone knows (that/if s.th. is the case//what is the case)>). The individual arguments, from which the structure patterns are composed, are explained with regard to their semantic role, their syntactic function, and the possibilities of morphosyntactic realization. The information in the transcript boxes illustrates the use of the patterns with a short transcript excerpt.

b. **Fixed phrases/collocations (7)**: Under this broad generic term, we subsume different types of more or less fixed lexical units (e.g. collocations, routine formula, proverbs) without further specification or terminological precision. These lexical units and collocations (e.g. *Bescheid wissen* [engl. *to be in the know*], *man weiß es ja nie* [engl. *you never know*]) are described, if considered relevant, in their semantic and/or formal properties and they are also individually illustrated with a transcript box.

c. The listing of **interactional units (8)**, which could be documented in the LeGeDe-sample in direct relation to specific meanings of certain lemmas (e.g. *keine Ahnung* [connection to *Ahnung* in the sense 'Wissen' [engl. 'knowledge'], *ich weiß nicht* [connection to *wissen* in the sense of 'to be informed']), enables a direct cross-connection to interactional functions (e.g. 'Unsicherheitsmarker': *ich weiß nicht* [engl. 'epistemic hedge': *I don't know*]), which are described in module 2.

(iv) **Other peculiarities (9)**: Furthermore, it is possible to point out interesting data in relation to selected metadata and their frequency.

Figure 5: Excerpt of *wissen⁸* ('informiert sein') [engl. *to know*, 'to be informed'] (Screenshot).

---

⁸ The author of the lexicographic article *wissen* (module 1: meanings) is Meike Meliss (member of the LeGeDe-Team).

5.2.3 Module 2: Functions in interaction

Module 2 describes the function of one- and multiword lemmas in spoken interaction (e.g. *eben, keine Ahnung, ich weiß nicht*). The different information items (1-7) will be explained using the example of the interactional unit *ich weiß nicht* [engl. *I don't know*] (cf. Fig. 6).

The general **cross-functional information (1)** is divided into categorical (modal particles (Thurmair, 1989), discourse particles (Willkop, 1988), etc.) and formal information regarding the elements involved in complex forms (e.g. *ich weiß nicht* [engl. *I don't know*]: verbal phrase; *keine Ahnung* [engl. *no idea*]: nominal phrase). Furthermore, formal information concerning the elements of multiword units (e.g. *ich weiß nicht*: "Phrase aus dem Personalpronomen *ich*, [...]" [engl. phrase formed from the personal pronoun [...]) and information on possible formal variants is offered on phonetic (e.g. *eben*: [ebent]; "Epithese eines stimmlosen [t]" [engl. "epithesis of an unvoiced [t]"), and compositional levels (e.g. *(ich) weiß nicht/weiß (ich) nicht*). A list of documented possibilities for combinatorics with an optional comment completes the general information together with a reference on the relevant research literature.

Each particular function that can be assigned to a lemma is **labelled (2)** accordingly. For example, *ich weiß nicht* as a multiword unit has a functional spectrum of different possibilities ('Unsicherheitsmarker', 'Markierung potenzieller Unangemessenheit'; [engl. 'epistemic hedge' or 'display of potential inappropriateness']). A **short description (3)** of the functions should help to differentiate the various possibilities. A **transcript box (4)** with the corresponding transcript excerpt, title, context, and comment is used for illustration.

In **abstraction of function (5)** generic information is offered. This information refers to findings which go beyond the occurrences in individual transcripts and therefore point at conspicuous features that have been revealed in the sample across the transcripts. These features are explained more in depth but in a comprehensive manner every user of each target group is able to grasp.

In addition to formal, categorical, combinatorial, and functional information, module 2 is enriched by information on **syntax and sequence realization (6)** and **prosody (7)** which are both illustrated by short transcript excerpts.

Figure 6: Excerpt of *ich weiß nicht*[9] ('Markierung potenzieller Unangemessenheit') [engl. *I don't know*, 'display of potential inappropriateness'] (Screenshot).

---

[9] The author of the lexicographic article *ich weiß nicht* (module 2: functions in interaction) is Katja Arens (a member of the LeGeDe-Team).

5.2.4 Links between module 1 and module 2

There is a very crucial connection between the two modules. In module 1, information about patterns and constructions is offered in the "combinatorics" section. Among these constructions are those with an underlying structural pattern and a syntactically functional approach. In addition, there are patterns or constructions in our data which have a special function in conversation. These "interactional units" are listed in module 1 in the section "combinatorics" (e.g. *ich weiß nicht* [engl. *I don't know*] as part of the dictionary article to *wissen* in the sense of 'informiert sein' [engl. 'to be informed']), but they are described in more detail in module 2. There are offered separate dictionary articles for the construction *ich weiß nicht* with a description of the function 'Unsicherheitsmarker' [engl. 'epistemic hedge'] (cf. 5.2.2., iii.c).

The semantic connection from module 2 to module 1 can also be illustrated by using the example of the multiword unit *ich weiß nicht* [engl. *I don't know*]. The short functional description in module 2 informs the user of the basic meaning contained in this pattern offering a reference to the sense 'informiert sein' [engl. 'to be informed'] of the verb *wissen* and links to module 1 accordingly (cf. Fig. 6).

## 5.3 Links within the resource: Outer texts

The dictionary user is offered four different types of outer texts. A section "About the LeGeDe-project" ("Über LeGeDe") provides a detailed reference about the project in general and to conceptual considerations about the LeGeDe-resource. In the "Usage instructions" section, a dictionary user learns in a guided tour how to navigate the resource and what types of information are offered. Very central terms used in our dictionary articles can be looked up in a "Glossary". Especially for grammatical terms, links to "grammis" (=Grammatisches Informationssystem) are offered via the glossary entries. Technical terms from the field of interactional linguistics – e.g. "Bezugsäußerung" [engl. *reference expression*], "Diskursmarker" [engl. *discourse marker*] or "Sequenz" [engl. *sequence*] – are explained and supplemented with research literature. From the glossary as well as from the dictionary articles we indicate very fundamental "Research literature". This can be viewed at a glance in a literature list.

## 5.4 Connection between the LeGeDe-resource and the DGD

A link to a lemma in the DGD database is offered in the overview article (cf. Fig. 3). Besides that, many details about a headword are supplemented in the dictionary article with authentic examples taken from the FOLK corpus (cf. Section 3). For each transcript excerpt there is the possibility to access the DGD database directly. For this purpose, it is necessary to create a personal account. After registering with the database, it is possible to view the transcript excerpts from the dictionary directly in the database, listen to the audio of the transcripts, sometimes even view video material and continue

researching the database. Via the overview article, the user is also able to search for a lemma in the *Lexical Explorer* (cf. section 3 and Batinić-Lemmenmeier in press).

## 6. Concluding remarks

As has been shown, the innovative aspects of the LeGeDe-resource are numerous. Since the project could not rely on any previous models, the simple fact of having created a lexicographic prototype to represent the specifics of the German spoken-language lexicon, using a corpus of spoken language in interaction as a basis, can be considered as a ground-breaking result. The conceptual considerations were based on assumptions from research, the LeGeDe-project work, and the results of the studies carried out during the LeGeDe-project. Concrete innovative aspects of the resource include the following:

(i)    Data basis: The resource is based exclusively on corpus-based data.

(ii)   Method: The corpus-based data have been quantitatively determined and qualitatively analysed and structured by a methodological approach developed by the team.

(iii)  List of headword candidates: The list of headword candidates was compiled using a specially developed corpus-based method of frequency comparison between two corpora: DeReKo as the reference corpus for written German and FOLK as the reference corpus of spoken interactional German.

(iv)   Range of information: The information offered on the lemmas is multimodular. The dictionary user finds a combination of traditional lexicographic information with an innovative offer of information which is developed specifically for the description of interactional functions. This is the first time that a proposal for new lexicographic information has been developed for the presentation of lexical phenomena of spoken language in interaction, which makes it possible to adequately structure and describe the specific phenomena for lexicographic purposes.

(v)    Authentic corpus evidence: Authentic corpus evidence is initially offered via selected transcript excerpts that provide an interface to the audio files and detailed information on the metadata. This makes the LeGeDe-resource one of the few lexicographic resources that has a direct, non-automatically generated link to the corresponding corpus data.

(vi)   Multimedia: The resource's multimedia character is characterized by the fact that, in addition to the transcripts, audio files and, in some cases, corresponding video files are available for the corpus data via access to the DGD. The link to the *Lexical Explorer* offers the possibility of extended analysis options.

(vii) Consideration of empirical expectations: The completely new conception of a lexicographic resource for the representation of linguistic specifics enabled the concrete consideration of certain empirically raised expectations of future users of such a new resource.

Not all aspects could be considered and implemented into the developed prototype during the project duration. Thus, there are certainly still many interesting possibilities for further research and development, for example in the area of the phenomenon classes (word formation, deixis, vagueness, etc.) or the access possibilities via an extended search in order to respond to the corresponding expectations of the participants in our surveys. Although the resource and the analyses are very detailed and complex, we hope that experts can take a mediating position in order to also make the contents accessible to different kinds of L1- and L2-learners.

# 7. References

Batinić-Lemmenmeier, D. (in press). Lexical Explorer: Extending access to the Database of Spoken German for user-specific purposes. In: *Corpora*, 15(1).

Bergmann, P. (2017). Gebrauchsprofile von *weiß nich* und *keine Ahnung* im Gespräch. Ein Blick auf nicht-responsive Vorkommen. In H. Blühdorn et al. (eds.), pp. 157–182.

Blühdorn, H., Deppermann, A., Helmer, H. & Spranz-Fogasy, T. (eds.) (2017). *Diskursmarker im Deutschen. Reflexionen und Analysen.* Göttingen: Verlag für Gesprächsforschung.

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

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.) *Duden – Die Grammatik.* Berlin: Dudenverlag, pp. 1181–1260.

Günthner, S. (2017). Diskursmarker in der Interaktion – Formen und Funktionen univerbierter *guck mal-* und *weißt du-*Konstruktionen. In H. Blühdorn et al. (eds.), pp. 103–130.

Handwerker, B., Bäuerle, R. & Sieberg, B. (eds.) (2016). *Gesprochene Fremdsprache Deutsch.* (Perspektiven Deutsch als Fremdsprache 32). Baltmannsweiler: Schneider.

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.* Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, pp. 929–935.

Helmer, H. & Deppermann, A. (2017). ICH WEIß NICHT zwischen Assertion und

Diskursmarker: Verwendungsspektren eines Ausdrucks und Überlegungen zu Kriterien für Diskursmarker. In H. Blühdorn et al. (eds.) *Diskursmarker im Deutschen. Reflexionen und Analysen.* Göttingen: Verlag für Gesprächsforschung, pp. 131–156.

Helmer, H., Deppermann, A. & Reineke, S. (2017). Antwort, epistemischer Marker oder Widerspruch? Sequenzielle, semantische und pragmatische Eigenschaften von *ich weiß nicht.* In A. Deppermann et al. (eds.) *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch.* (Studien zur deutschen Sprache 74). Tübingen: Narr, pp. 377–406.

Imo, W. (2007). Zur Anwendung der Construction Grammar auf die gesprochene Sprache – Der Fall *„ich mein(e)“.* In V. Ágel & M. Hennig (eds.) *Zugänge zur Grammatik der gesprochenen Sprache.* (Germanistische Linguistik 269). Tübingen: Niemeyer, pp. 3–34.

Imo, W. & Moraldo, S. M. (eds.) (2015). *Interaktionale Sprache und ihre Didaktisierung im DaF-Unterricht.* (Deutschdidaktik 4). Tübingen: Stauffenburg.

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

Klosa, A. & Müller-Spitzer, C. (eds.) (2016). *Internetlexikografie. Ein Kompendium.* Berlin/Boston: de Gruyter.

Klosa, A. 2013. 'The Lexicographical Process (with Special Focus on Online Dictionaries)'. In R. H. Gouws et al. (eds.) *Dictionaries. An international encyclopedia of lexikography. Supplementary volume: Recent developments with focus on electronic and computational lexicography.* (Handbücher zur Sprach- und Kommunikationswissenschaft 5). Berlin: de Gruyter, pp. 517–524.

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.

Kupietz, M. & Schmidt, T. (2015). Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In L. M. Eichinger (ed.) *Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven.* (Jahrbuch des Instituts für Deutsche Sprache 2014). Berlin: de Gruyter, pp. 297–322.

Kupietz, M., Lüngen, H., Kamocki, P. & Witt, A. (2018). The German Reference Corpus DEREKO: New Developments – New Opportunities. In N. Calzolari et al. (eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki: European Language Resources Association (ELRA), 2018. pp. 4353–4360.

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

Meliss, M. & Möhrs, C. (2017). Die Entwicklung einer lexikografischen Ressource im Rahmen des Projektes LeGeDe. *Sprachreport 4/2017*, pp. 42–52.

Meliss, M. & Möhrs, C. (2018). Lexik in der spontanen, gesprochensprachlichen Interaktion: Eine anwendungsorientierte Annäherung aus der DaF-Perspektive. *GFL*, 3/2018, pp. 79–110.

Meliss, M., Möhrs, C., Batinić, D. & Perkuhn, R. (2018). Creating a List of Headwords for a Lexical Resource of Spoken German. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (eds.) *Proceedings of the 18th EURALEX International Congress 2018. 17-21 July 2018.* Ljubljana, Slovenia, pp. 1009–1016.

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. *Zeitschrift für Angewandte Linguistik*, 68(1), pp. 103–138.

Meliss, M., Möhrs, C. & Ribeiro Silveira M. (2019). Anforderungen und Erwartungen an eine lexikografische Ressource des gesprochenen Deutsch aus der L2-Lernerperspektive. *Lexicographica*, 34, pp. 89–121.

Meliss, M., Möhrs, C. & 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. 19-21 September 2017. Leiden, the Netherlands.* Brno: Lexical Computing CZ s.r.o., pp. 281–298.

Moon, R. (1998). On Using Spoken Data in Corpus Lexicography. In T. Fontenelle, P. Hiligsmann, A. Michiels, A. Moulin & S. Theissen (eds.) *Proceedings of the 8th EURALEX International Congress. 4-8 August 1998. Liège, Belgium.* Liège: English and Dutch Departments, University of Liège, pp. 347–355.

Moraldo, S. M. & Missaglia, F. (eds.) (2013). *Gesprochene Sprache im DaF-Unterricht. Grundlagen – Ansätze – Praxis.* (Sprache – Literatur und Geschichte 43). Heidelberg: Winter.

Reeg, U., Gallo, P. & Moraldo, S. M. (eds.) (2012). *Gesprochene Sprache im DaF-Unterricht. Zur Theorie und Praxis eines Lerngegenstandes.* (Interkulturelle Perspektiven in der Sprachwissenschaft und ihrer Didaktik 3). Berlin: Waxmann.

Schmidt, T. (2014a). 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 International Conference on Language Resources and Evaluation. 26-31 May 2014. Reykjavik, Iceland.* Reykjavik: European Language Resources Association, pp. 383–387.

Schmidt, T. (2014b). The Database for Spoken German – DGD2. In N. Calzolari et al. (eds.) *Proceedings of the 9th International Conference on Language Resources and Evaluation. 26-31 May 2014. Reykjavik, Iceland.* Reykjavik: European Language Resources Association, pp. 1451–1457.

Schmidt, T. (2017). DGD – Die Datenbank für Gesprochenes Deutsch. Mündliche Korpora am Institut für Deutsche Sprache (IDS) in Mannheim. *Zeitschrift für Germanistische Linguistik*, 45(3), pp. 451–463.

Schwitalla, J. (⁴2012). *Gesprochenes Deutsch. Eine Einführung.* (Grundlagen der Germanistik 33). Berlin: Schmidt.

Sieberg, B. (2013). *Sprechen Lehren, Lernen und Verstehen. Stufenübergreifendes Studien- und Übungsbuch für den DaF-Bereich.* Tübingen: Julius Groos.

Siepmann, D. (2015). Dictionaries and Spoken Language: A Corpus-Based Review of French Dictionaries. *International Journal of Lexicography*, 28(2), pp. 139–168.

Thurmair, M. (1989). *Modalpartikeln und ihre Kombinationen.* Tübingen: Niemeyer.

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.* Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud, pp. 311–318.

Trim, J., North, B., Coste, D. & Sheils, J. (2001). *Gemeinsamer Europäischer Referenzrahmen für Sprachen: Lernen, Lehren, Beurteilen. Niveau A1, A2, B1, B2, C1, C2.* Berlin: Langenscheidt.

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

Willkop, E.-M. (1988). *Gliederungspartikeln im Dialog.* München: Iudicium.

Zeschel, A. (2017). *Denken* und *wissen* im gesprochenen Deutsch. In A. Deppermann, N. Proske & A. Zeschel (eds.) *Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch.* (Studien zur Deutschen Sprache 74). Tübingen: Narr, pp. 249–336.

**Websites (5 August 2019)**

*DeReKo.* Accessed at: http://www1.ids-mannheim.de/kl/projekte/korpora.html.

*DGD.* Accessed at: https://dgd.ids-mannheim.de.

*DWDS.* Accessed at: https://www.dwds.de/.

*FOLK.* Accessed at: http://agd.ids-mannheim.de/folk.shtml.

*grammis.* Accessed at: https://grammis.ids-mannheim.de/.

*LeGeDe project website:* http://www1.ids-mannheim.de/lexik/lexik-des-gesprochenen-deutsch.html.

*LeGeDe-Resource*: Accessed at: https://www.owid.de/legede/ (available from September 2019).

*Lexical Explorer.* Accessed at: https://www.owid.de/lexex/.