# A corpus-based dictionary of spoken German – a short insight into a new project

## Christine Möhrs

Institut für Deutsche Sprache, P.O. Box 101621, D-68016 Mannheim
E-mail: moehrs@ids-mannheim.de

## Abstract

This paper presents a short insight into a new project at the "Institute for the German Language" (IDS) (Mannheim). It gives an insight into some basic ideas for a corpus-based dictionary of spoken German, which will be developed and compiled by the new project "The Lexicon of spoken German" (Lexik des gesprochenen Deutsch, LeGeDe). The work is based on the "Research and Teaching Corpus of Spoken German" (Forschungs- und Lehrkorpus Gesprochenes Deutsch, FOLK), which is implemented in the "Database for Spoken German" (Datenbank für Gesprochenes Deutsch, DGD). Both resources, the database and the corpus, have been developed at the IDS.
**Keywords:** electronic lexicography; spoken German; project report

## 1. Introduction

In this paper we want to present a short insight into a new project, started in September 2016 at the "Institute for the German Language" (IDS). The project is called "Lexik des gesprochenen Deutsch"[1] (eng.: "The Lexicon of spoken German"), abbreviated LeDeGe in the following. The goal of the project is to develop a dictionary of spoken German by using corpus-based methods. The work is based on the "Forschungs- und Lehrkorpus Gesprochenes Deutsch, FOLK" (eng.: "Research and Teaching Corpus of Spoken German"), which is implemented in the "Datenbank für Gesprochenes Deutsch, DGD" (eng.: "Database for Spoken German"). Both resources, the database and the corpus, have been developed at the IDS. In section 2 we want to give a short insight into the corpus database of the project.

## 2. The project LeDeGe

The project LeGeDe is a third-party funded project of the Leibniz Association (Leibniz Competition 2016, Funding line 1: Innovative projects[2]). During the next three years (2016-2019) the project will be working on the innovative idea of a dictionary of spoken German. Such a dictionary resource does not yet exist, as modern dictionaries of German are usually based on the written language that is represented in large

---

[1] For an overview about the project, please see http://www1.ids-mannheim.de/lexik/lexik-des-gesprochenen-deutsch.html.
[2] For more information about the competition and the funded projects, please go to: http://www.leibniz-gemeinschaft.de/en/about-us/leibniz-competition/projekte-2016/funding-line-1/.

electronic text corpora. As a prerequisite, lexical phenomena that are typical for spoken German are analyzed on the basis of the corpora of spoken German available at the "Institute for the German Language".

The project is to be achieved through a cooperation of two departments of the "Institute for the German Language" in Mannheim: the Department of Pragmatics and the Department of Lexical Studies. The team consists of researchers with different research areas: lexicographers (especially researchers with s special focus on electronic lexicography), corpus linguists, and researchers with a special focus on conversational analysis[3].

# 3. Database

FOLK – the research and teaching corpus of spoken German – is a corpus project, which is located in the Department of Pragmatics. As described in Schmidt 2014a:

"The project 'Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)' (eng.: research and teaching corpus of spoken German) has […] set itself the aim of building a corpus of German conversations which:

a) covers a broad range of interaction types in private, institutional and public settings,

b) is sufficiently large and diverse and of sufficient quality to support different qualitative and quantitative research approaches,

c) is transcribed, annotated and made accessible according to current technological standards,

d) is available to the scientific community on a sound legal basis and without unnecessary restrictions of usage." (Schmidt 2014a: 383; cf. Fig. 1)

---

[3] LeGeDe-Team: Katja Arens, Dolores Batinić, Arnulf Deppermann, Stefan Engelberg, Meike Meliss, Christine Möhrs, Thomas Schmidt, Antje Töpel, Sarah Torres.

Figure 1: Extract from search results to *gucken* (FOLK, DGD)

By today, a set of data comprising over 170h of recordings and close to 1,600,000 transcribed tokens has been completely processed and published via the "Database for Spoken German" (cf. Schmidt 2014b). During the project period we want to develop methods and a lexicographical process that take the characteristics of spoken language and the possibilities of the database into account. In addition to that we want to differentiate between aspects of spoken and written language.

## 4. Goals

The aim of the project "The Lexicon of Spoken German" is to design and compile the first dictionary of spoken German. The dictionary will be available online and will be extensible. It will cover the lexical units and properties typical for spoken German as it is used in conversations in private and institutional contexts. In the future, the dictionary will be integrated into the dictionary portal OWID [4] (Online-Wortschatz-Informationssystem Deutsch; eng.: Online vocabulary system of the German language).

---

# 5. References

**Paper in conference proceedings:**

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.

**Websites:**

*Datenbank zum Gesprochenen Deutsch.* Accessed at: http://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome. (14 December 2016)

*OWID.* Accessed at: www.owid.de. (14 December 2016)