PhonD2: A DATABASE ON PHONOTACTIC STRUCTURES OF GERMAN DIALECTS

Alfred Lameli[1], Alexander Werth[2], Valeria Bunkov[2], Samantha Link[1]

Research Center Deutscher Sprachatlas, Philipps-Universität Marburg, Germany[1]
University of Passau, Germany[2]
lameli@uni-marburg,de, alexander.werth@uni-passau.de, valeria.bunkov@uni-passau.de, samantha.link@uni-marburg.de

ABSTRACT

We present the PhonD2 corpus, which is an open access online database on phonotactic structures of German dialects. The corpus is based on both the translation of sentences and free speech during interviews in 172 sites all over Germany. Data are annotated according to phonological and morphological criteria, e.g., sounds, CV-structure, morphemes. At present, the sub-corpus of translations is integrated, free speech will be added successively. The corpus focuses on syllable structures in dialects aiming at the description of geographical variation of syllable patterns at the phonology-phonetics interface and their typological classification. It shows, e.g., systematic regional differences in preferences for monosyllabic words, differences in how they arrange on the sonority scale as well as regional differences in the clustering of consonants in both onset and coda. The PhonD2 corpus thus opens up a systematic perspective on the clustering of sounds and on typological characteristics of German beyond standard language.

Keywords: Phonotactics, German dialects, syllable, sonority, CV phonology.

1. INTRODUCTION

Since the earliest days of dialectology, there has been an awareness of the fact that the realization of sounds may differ depending on their linguistic context. Such phonotactic conditions have always been taken into account in the relevant grammars of local dialects (e.g., [1]). However, they have not been systematically researched on a comparative regional scale. This is mainly due to the lack of data based on spoken dialects and suited for phonotactic questions. Although numerous corpora of spoken German language have been compiled in recent years (cf. [2]), they are either not regionally diversified or, if they are, they are not phonetically-phonologically specified.

This paper describes the planning and present implementation of the PhonD2 ("Phonotaktik der Dialekte in Deutschland") database, which is a database at phonotactic structures of German dialects. The focus of the PhonD2 database is on the regional documentation of syllable structures in Germany. For this purpose, data from both the translation of sentences and free speech during interviews documented in the *Marburg Phonetic Archive* (MRPhA) were selected from 172 sites all over Germany. First analyses based on this data indicate large-scale differences in the preference of syllable structures between dialect regions, which can be interpreted in terms of (micro)typological variation [3].

The aim of this paper is to introduce the PhonD2 database and to illustrate its linguistic potential. The remainder of the paper is as follows. Section 2 provides information on the PhonD2 data and its processing, section 3 takes a closer look on the data and section 4 concludes.

2. DATA AND DATA PROCESSING

2.1 Background

The data used for the PhonD2 database were selected from the *Marburg Phonetic Archive* (MRPhA), which is a largescale archive of dialectal sound recordings ranging from the first half of the 20th century to the present. The archive is maintained by the *Research Center Deutscher Sprachatlas* (DSA, Marburg/Germany).

Initially, the PhonD2 project made use of the MRPhA selection chosen by Göschel and colleagues in preparation of the *Phonetischer Atlas der Bundesrepublik Deutschland* (PAD; cf. [4, 5]). This selection documents the speech of so-called NORMs and NORFs (non-mobile, old, rural (fe)males), which were recorded between 1956 and 1996 by phonetically trained DSA explorers. The data collection was aimed at determining the oldest dialect forms known by these persons. For this purpose, translation tasks of Wenker sentences were carried out. The Wenker sentences are a standard instrument in German dialectology [6] consisting of 40 sentences plus additional words (e.g., numbers, weekdays).

Göschel selected a total of 184 recordings each of which was transcribed in parts during the 1980s and 1990s in a very narrow IPA notation supported by quantifying control measures to ensure the highest possible transcription quality (cf. [7]). Even though these transcriptions have been completed, the atlas remained unpublished. The data was released to the public for further analyses and have been used several times in dialectometric studies since then (e.g., [8, 9, 10, 11, 12]).

The PhonD2 corpus makes use of this PAD selection and its transcriptions of Wenker sentences. In order to be able to consider a further speech situation, additional recordings of the same persons from the MRPhA were integrated into the corpus. These are interviews on biographically relevant topics that were also recorded during the original data exploration. The monologues that developed from this, interrupted by short queries, form an excellent documentation of the free speech of these persons. However, since the documentation of free speech is not available for all PAD informants, the number of sites was reduced for the PhonD2 project. The PhonD2 corpus thus contains a total of ca. 20 hours of spoken language from 172 sites.

2.2 Workflow

In order to compute both a phonetic segmentation and labeling of the recordings WebMAUS [13] is used. We use WebMAUS' standard German language model, but enhance the orthographic notation required for it with dialectal phenomena and phenomena of spoken language such as clitization or sentence and word interruptions. This approach significantly improves the alignment of the dialect data. Further steps are the production of both a phonological and morphological representation as well as the implementation of language data and meta-data in the PhonD2 database.

2.2.1 Phonological representation

In order to prepare the phonotactic documentation, the PAD transcriptions must be modified. Take, for example, the notation [ˈmɔɪ̯ɪ̯də] 'tired' from the PAD corpus. Two essential work steps are involved here, namely the normalization of the transcript and the identification of syllables. The IPA transcriptions are thus assigned to phonological types that could be derived from the research literature. The identification of syllables is processed by a rule-based algorithm, the results of which are nevertheless checked by visual inspection. In the given case this procedure results in the representation [ˈmɔɪ̯.də].

The parts of the Wenker sentences, which have not been prepared during the PAD work, were directly transcribed. The same holds for the data from the situation of free speech context. Since some dialects are difficult to understand, we use a network of native dialect speakers to assist us in preparing the orthographic transcripts. We also do not transcribe the entire monologues, but only individual words, usually nouns, verbs, adjectives and adverbs.

The basis for the comparability of the transcriptions is, on the one hand, continuous training of the transcribers (including the comparison and control of individual transcriptions), and, on the other hand, a constantly growing catalog of regional features compiled in the course of the project.

Using these notations, a phonological representation that distinguishes plosives (P), affricates (A), fricatives (F), nasals (N), liquids (L), glides (G), short vowels (V), long vowels (V:) and diphthongs (VV) is automatically derived and implemented into the PhonD2 database (e.g., [ˈmɔɪ̯.də] ~ NVV.PV) as well as a less detailed CV representation ([ˈmɔɪ̯.də] ~ CVV.CV).[1] Furthermore, an indication of sonority with 1 = vowel, 2 = glide, 3 = liquid, 4 = nasal, 5 = fricative, 6 = plosive is implemented ([ˈmɔɪ̯.də] ~ 51.61) together with the information on strong and weak syllables ([ˈmɔɪ̯.də] ~ s.w). In addition, primary word accents and Middle-Franconian tone accents are annotated where necessary.

The classifications are sometimes difficult to implement, for example, with regard to the question as to whether a certain sound has to be defined as a glide or a fricative. From this point of view, the classification is to be understood as a suggestion (based on reference literature such as [14]), which can be modified in the context of individual data processing.

2.2.2 Morphological representation

In an earlier study on monosyllables [12], a dependency between PoS and phonotactic structure has been found for German dialects, in that some PoS do systematically prefer or disfavor certain syllable structures. Consequently, the PhonD2 data allows to systematically explore the interface between phonotactics and morphology. Based on STTS tagset [15], we automatically assigned PoS using TreeTagger [16]. To avoid inaccuracies, the tagged PoS are corrected manually. Tags not available in STTS are individually assigned based on the relevant literature on speech classifications for German (e.g., [17]). Usually, a word form is assigned to exactly one PoS; but some contexts require a word form assignment to more than one PoS (e.g., *mit* 'with' as a preposition vs. *mit* 'with' as a particle as in *mitgehen* 'to go along'). In such cases multiple individual tags

were provided (e.g., *mit* is assigned to "APPR" and "PTKVZ"). Apart from PoS, morphemes of the German dialects are segmented and identified manually. The corpus scheme for morphemes consists of (word-)stem, affixes, vowel mutation and vowel gradation, diminutives and suppletive forms.

2.3 Online access

All data will be available open access on the Internet.[2] At present (01/2023), the PhonD2 site provides an overview of several items of the Wenker sentences together with a map of the survey locations. Other data will be integrated successively. In addition, a mapping function illustrates the regional distribution of syllable types, for which Figure 2 and Figure 3 provide examples. These maps are intended only as a small glimpse into the geographical dimension of the data. For exhaustive analyses, the data must be processed individually. We do not achieve full coverage of the 172 sites in all cases; these instances are nevertheless included, provided they can be analyzed in phonotactic terms.

Additionally, the web site provides a first overview on sound frequencies regarding both individual sounds and sound combinations. In this regard, Figure 1 gives an impression of tri-grams with a liquid [l] core together with information on their frequency in the data.

trigrams	freq
blɐ	1
blɑ	1
ble	42
blə	2
blɛ	106
bli	62
blɪ	140
bln	6
blø	1
blœ	2

Figure 1: Extract of tri-grams with a liquid [l] core in the PhonD2 tri-gram table.

After completion of the project, the data can be downloaded in its entirety as a *.csv file. Until then, all data presented on the Internet must be considered preliminary. Nevertheless, they can be used for linguistic analyses.

3. DATA INSPECTION

3.1 Cartographic representations

Figure 2 illustrates the geographical distribution of syllable types of the standard German noun *Affe* 'ape' in the PhonD2 database. The map shows clear regional patterns in both the distribution of monosyllables vs. bisyllables (e.g., VC vs. V.CV ~ [af] vs. [ˈa.fə]) and the organization of syllables within those types (e.g., VC vs. V:C ~ [af] vs. [a:p]; glottal stops are not reported). It appears that monosyllables are preferred in the North and in the lower half of Germany, while bisyllabic realizations occur almost exclusively in a central strip stretching from West to East.

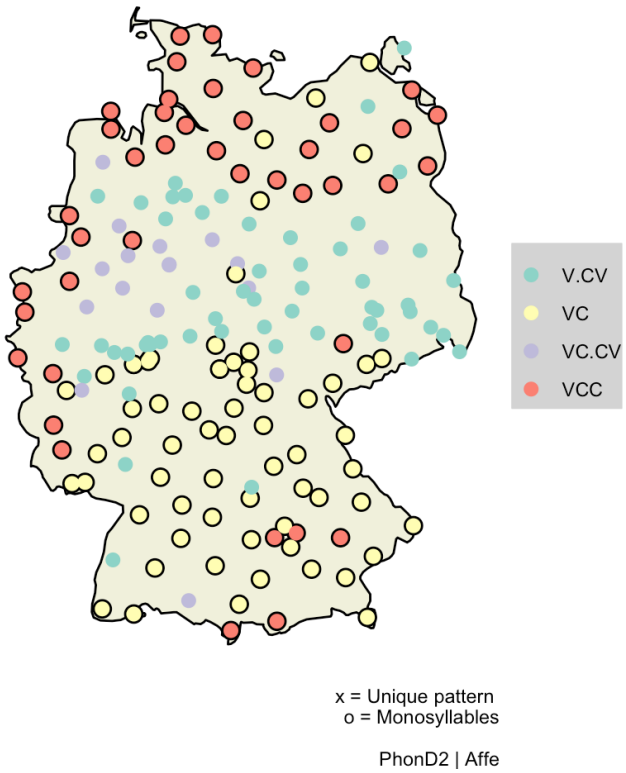

Figure 2: Regional distribution of syllable structure for the word *Affe* 'ape' from the PhonD2 corpus.

In Figure 2, several phonological processes come together, namely schwa apocopation ([af] vs. [ˈa.fə]), which indicates preference for monosyllabic or bisyllabic syllables and monosyllabic lengthening ([af] vs. [a:p]) [18, 19]. Against this background, the phonotactic map provides a linguistic systematization on a higher level; it classifies linguistic patterns with respect to more general structural characteristics.

Figure 3 represents the standard German adverb *genug* 'enough'. Again, there are preferences for monosyllabic forms in the North and in the South, even though the zone with bisyllabic forms between them is much more expanded. In this case, however, monosyllables are not attributed to schwa

apocopation, but schwa syncopation (e.g., [ˈknʊɐ̯] ~ CCVV or [ˈknʊːx] ~ CCV:C) or the loss of the *ge-*syllable ([ˈnɔx] ~ CVC or [ˈnaʊ̯x] ~ CVVC). The latter is interesting from the perspective that *ge-* is usually used as a prefix in German (e.g., past participle *gekannt* 'known'), which typically is lost in, for example, Bavarian dialects (*Ø-kent*). In the adverb *genug*, *ge-* is, however, no morphological marker thus indicating a phonological process driven by analogy. Regarding the distribution of syllable types, the pattern known from the *Affe* map is mirrored here to some extent.

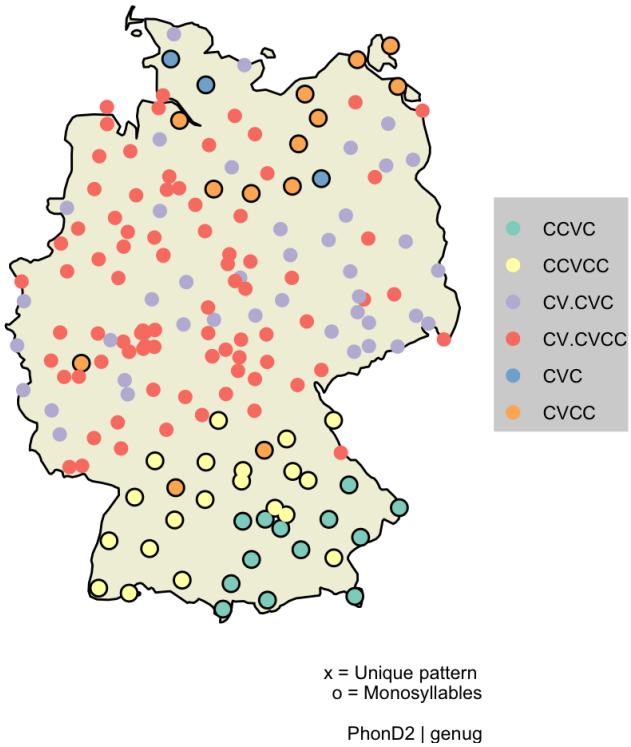

Figure 3: Regional distribution of syllable structure for the word *genug* 'enough' from the PhonD2 corpus.

At the same time, Figure 3 shows one of the rare examples, where onset and coda of a word are very closely arranged on the sonority scale as in [ˈgnʊŋk] ~ PNVNP or [ˈgnuːŋk] ~ PNV:NP [14]. With this pattern in mind, schwa syncopation can be understood as an optimization of syllable structure toward the most consistent sonority arrangement in the onset, supported by the integration of a velar nasal in the coda. From this point of view, we are not only dealing with the sonority optimization of the syllable, but the sonority optimization of the (phonological) word. As can be seen, this process is restricted to the Upper German dialect region (= yellow and green color). Whether this is a typological characteristic of Upper German dialects must be left to further analysis.

3.2 Data aggregation

In addition to single word representations, the aggregate evaluation of phonotactic structures is of interest. Figure 4 shows Finite State Automata (Markov Chains) for three subsamples of the PhonD2 corpus ([3]), representing the transition probabilities between consonants (C) and vowels (V) in monosyllables.

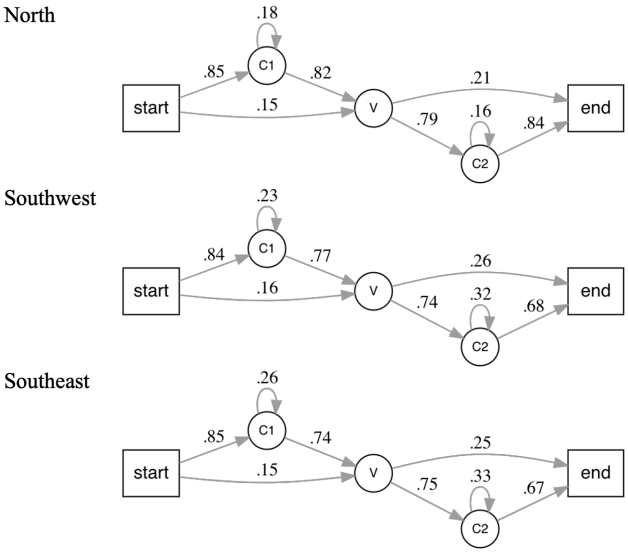

Figure 4: Finite State Automata for regional PhonD2 subsamples documenting CV transition probabilities in monosyllables

The Figure shows that in all regions, the probability that a monosyllable starts on C is about 85 %. However, then regional differences in the replication rate of the C position in Onset (C1) become apparent: 18 % in the North, 23%, 26 % in the South. It is similar for the replication rate in the coda (C2): 16 % in the North, 32 %, 33 % in the South. What the PhonD2 data here more generally shows is a regional difference in the syllable complexity of German dialects in the form of a South-North divide: more C clusters in the South, fewer C clusters in the North.

4. SUMMARY

We report on the PhonD2 database, which is a database on syllable structures of German Dialects. The phonotactic account enables a comprehensive view on more general structural characteristics of dialects, both with regard to individual phonetic characteristics, but also with regard to the typological structure of the dialects. The database is constantly being expanded. All data are freely accessible.

Acknowledgements
This research is funded by the German Research Foundation (DFG, grant 432304149). We are grateful

to two anonymous reviewers and to the discussants at project presentations in Marburg, Frankfurt, Vienna, and Salzburg. Thanks also go to both our student workers and our transcribers in the language regions.

5. REFERENCES

[1] Winteler, J. 1876. *Die Kerenzer Mundart des Kantons Glarus. In ihren Grundzügen dargestellt*. Winter.
[2] Kupietz, M., Schmidt, T. (eds) 2018. *Korpuslinguistik*. De Gruyter.
[3] Lameli, A. 2022. Syllable Structure Spatially Distributed: Patterns of Monosyllables in German Dialects. *Journal of Germanic Linguistics* 34, 241–287.
[4] Göschel, J. 1992. *Das Forschungsinstitut für Deutsche Sprache "Deutscher Sprachatlas"*. Forschungsinstitut für deutsche Sprache.
[5] Göschel, J. 2000. Der Phonetische Atlas von Deutschland. *Јужнословенски Филолог* 56, 283–288.
[6] Chambers, J. K., Trudgill, P. 1998. *Dialectology. Second edition*. CUP.
[7] Almeida, A., Braun, A. 1986. "Richtig" und "Falsch" in phonetischer Transkription. Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. *Zeitschrift für Dialektologie und Linguistik* 53, 158–172.
[8] Nerbonne, J. Siedle, C. 2005. Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik* 72, 129–147.
[9] Nerbonne, J. 2010. Mapping aggregate variation. In: Lameli, A., Kehrein, R., Rabanus, S. (eds), *Language and space. An international handbook of linguistic variation, vol. 2: Language mapping, part I*. De Gruyter Mouton, 476–495.
[10] Prokić, J, Çöltekin, C., Nerbonne, J. 2012. Detecting shibboleths. In: Butt, M., Carpendale, S., Penn, G., Prokić, J., Cysouw, M. (eds), *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH: Visualization of linguistic patterns and uncovering language history from multilingual resources*. The Association for Computational Linguistics, 72–80.
[11] Prokić, J. 2017. Quantitative diachronic dialectology. In: Wieling, M., Kroon, M., van Noord, G., Bouma, G. (eds), *From semantics to dialectometry. Festschrift in honor of John Nerbonne*. College Publications, 293–301.
[12] Lameli, A., Werth, A. 2017. Komplexität und Indexikalität. Zum funktionalen Gehalt phonotaktischer Wortstrukturen im Deutschen. Hennig, M. (eds), *Linguistische Komplexität – ein Phantom?* Stauffenburg, 73–96.
[13] Kisler, T., Reichel U. D., Schiel, F. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45, 326–347.
[14] Hall, T. A. 1992. *Syllable structure and syllable-related processes in German*. Niemeyer.
[15] Schiller, A., Teufel, S., Stöckert, C. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset)*. University of Tuebingen.
[16] Schmid, Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.
[17] Hentschel, E., Weydt, H. 2021. *Handbuch der deutschen Grammatik*. de Gruyter. 5. Ed.
[18] Schirmunski, V. 2010 [1962]. *Deutsche Mundartkunde. Vergleichende Laut- und Formenlehre der deutschen Mundarten*. Lang.
[19] Lameli, A. 2022. Remarks on the consistency of schwa apocope in the geography of German dialects. In: Nevaci, M., Floarea, I, Farcaş, J-M (eds), *Ex Oriente lux. In honorem Nicolae Saramandu*. Edizioni dell'Orso. 683–702.

[1] We differentiate between V: and VV in order to preserve the phonetic basis. Of course, unification can be done post hoc in the sense of CV representation (V:, VV = VC).
[2] http://www.dsa.info/phond2.