Ralf Plate

WORD FAMILIES IN DIACHRONY

An epoch-spanning structure for the word families of older German

Abstract The 'Word Families in Diachrony' project (WoDia), for which a funding application to the DFG is in preparation, aims to provide a database-driven online research environment that will enable processes of change in the entire historical vocabulary of German to be investigated by focusing on the changes in word families and the individual means of word formation. WoDia will embed the vocabularies of Old High German (OHG), Middle High German (MHG), Old Saxon (OS), and Middle Low German (MLG) in a database, resulting in a word-family structure for High and Low German from the beginnings up to the 15th century (for High German) and up to the 17th century (for Low German). The basis of the vocabulary is provided by reference dictionaries of the four historical varieties, whereas the word families' historical structure is based on the word-family dictionary of OHG by Jochen Splett (1992). Each lemma in the database will be assigned, where appropriate, to a word family. The individual word-formation elements and the word-formation hierarchy will be mapped in a structural formula. The etymologically corresponding lemmas and word families of the different periods/varieties of older German will be linked so that an analysis across the varieties will also be possible. The annotations of word families in the database (e. g., relating to word structure) will be supplemented by linking their lemmas to the online dictionaries and to the reference corpora of Old German (OS and OHG), MHG, and MLG.

Keywords Older German (OHG, MHG, OS, MLG); word family database; historical word formation of German

Introduction

The project presented here will address several desiderata of historical lexicography and historical word-formation research in German:

- The linking of the vocabularies of the older stages of High and Low German (OHG, MHG, OS, MLG), as documented in the reference dictionaries of these varieties, so that each word of each of the varieties is linked to its etymological counterparts in the others (insofar as they are extant).
- The morphological analysis of these vocabularies' word formation, so that for each word, its word-formation structure is recorded in a hierarchical formula.
- The structuring of these vocabularies into word families and the establishment of an overarching word-family structure, in which the individual vocabularies of the four varieties are brought together.

The results of the project will be made available online in WoDia (Word Families in Diachrony), a database-driven research environment with a user-friendly front end for research and teaching.

The project is able to draw on extensive preliminary work and resources, which are described under 1. The basis of the data and the necessary stages of work on the project are presented under 2.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

1. Preliminary work and resources

1.1 Jochen Splett's word-family dictionaries of OHG (Splett 1993) and of contemporary German (Splett 2009)

The reference dictionaries for the above-mentioned varieties of older German are based on single words and do not contain any systematic morphological information about the word formation of the lemmas. However, a comprehensive analysis of word-formation morphology has been carried out by Jochen Splett in two word-family dictionaries: on OHG (1992, 3 vols) and on contemporary German (2009, 18 vols). The particular distinction of these dictionaries, aside from the division of lexemes into word families, is that they offer a hierarchical (bracketed) structural formula for each lemma of a word family, which indicates the constituent structure and the part of speech of the stems. Cf. the following examples from the dictionary of contemporary language (vol. 1, pp. 416f.; cf. also Fig. 1):

- *Bau-er*: (wV)sS – To be read as: substantival derivative from the verbal stem BAU with the suffix *-er* (w = root / stem; V = verb; s = suffix; S = substantive).
- *Acker-bau-er*: (wS) ((wV)sS) – Composite with BAU-ER as the basic component and the stem ACKER as the determinative element.
- *Boots-bau-er*: ((wS) ((wV)S))sS – substantival derivation from the compound noun BOOTS-BAU with the suffix *-er*; *-bau* is again a substantival conversion of the verbal stem BAU.
- *Er-bau-er*: (p(wV))sS – noun derivation from the prefixed verb stem ER-BAU with the suffix *-er* (p = prefix).

The ambiguity arising from double or multiple motivations is systematically taken into account by specifying alternative structural formulae, e. g., in *baukünstlerisch* adj. (vol. 1, p. 414) with three formulae for the three possible analyses (*baukünstlerisch* or *bau-künstlerisch* as immediate constituents and again in the first case *bau-künstler* or *baukunst-ler* as immediate constituents of the first element).

In particular, the structural formulae, with their formalized description of the steps and word-formation elements on which the formations are based, offer a wide range of insights for questions of historical word formation theory. However, since Splett's dictionaries are not yet digitally accessible, their usefulness for such questions has so far been significantly restricted.

1.2 Thomas Klein's semi-automatic word-formation analysis and determination of word family stems

Apart from OHG, the material basis for comprehensive studies of historical word formation, the development and restructuring of vocabulary, and the means of word formation in general is still lacking. This needs to be developed through a morphological analysis of word formation for the historical vocabularies as in Splett's dictionaries. Thomas Klein has taken the initiative in addressing this desideratum, by presenting a methodologically convincing approach to a semi-automatic morphological analysis of word formation and word family classification, in his case, for Middle High German (most recently, Klein 2018). Klein's basic idea consists in the automatic segmentation of lemmas into affixes and stems and the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

automatic tracing back of the stems to that stem variant which appears in the root (core) word of their word family, the word family stem (WFSt), i. e., the form without OHG *i*-umlaut, ablaut, grammatical change, gemination, Germanic *i*- and *a*-umlaut, etc. In this way, for example, the stems (stem variants) GOLT and GÜLT are isolated from the MHG lexemes *unvergolten* (part. adj.) and *gültic* (adj.), by separating the affixes; they are then traced back to the WFSt GELT. These two and all other words of the MHG word family GELTEN then (ideally) appear under the WFSt GELT.

1.3 Testing of the transepochal word family linking in the ZHistLex Project

Whereas Klein (2018) focuses on the task of automatically determining the MHG WFSt, the further step of semi-automatically mapping the MHG vocabulary to the OHG word families was tested in a pilot project at the University of Frankfurt/Main (a part-project in the preparation of an eHumanties Centre for Historical Lexicography, ZHistLex, funded by the BMBF 2016–2019; cf. Plate 2020).

Exact OHG-MHG word equivalents exist for only one sixth (17%) of the MHG vocabulary; however, for the semi-automatic assignment of the remaining MHG vocabulary to OHG word families, the linking of their stems resulting from the linking of word equivalents can be used, as indicated in the above example (1.2).

Because MHG *unvergolten* was linked to earlier OHG *unfirgoltan* and thus also to its word family GELTAN, its WFSt GELT is also linked to the OHG word family GELTAN. The established correspondences of the type 'MHG WFSt GELT = OHG word family GELTAN' can now be used to assign MHG words automatically to an OHG word family without an OHG precursor if their WFSt (or in the case of compounds, at least one of the word family stems) is already recorded and assigned. Thus, the MHG lexeme *bete-gültic* 'liable to tax' (for which, as well as for its adjectival second component *gültic*, no OHG antecedent is attested) can be automatically assigned to the OHG word family GELTAN on the basis of the MHG WFSt GELT that had already been determined. The same applies to the first component *bete* with the WFSt *bit* and the OHG word family BITTEN. – The automatic assignment gives rise to lists of suggestions, which have to be carefully examined individually.

1.4 Online dictionaries and lemma lists

Lemma lists are a prerequisite for processing the vocabularies of the varieties of older German. These are extracted from the digitized reference dictionaries. In the case of Old High German and Middle High German, the dictionaries and lemma lists are already available digitally.

Reference dictionaries for Old High German are the *Althochdeutsches Wörterbuch* of the Leipzig Academy of Sciences (AWB), which is still being compiled and is available online in the Trier Dictionary Network, and, until the AWB is complete, the *Althochdeutsches Wörterbuch* by Jochen Splett (1992 = AWB^{Spl}), which covers the entire vocabulary of Old High German (approx. 28,500 words). It has been digitized at the University of Frankfurt and will also be published online in the Trier Dictionary Network as part of the project. A sample of the online publication in preparation is shown in Figure 1 on the following page; in the left column the list of word families and lemmas is shown, in the centre column the dictionary,

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

with links to the Leipzig *Althochdeutsches Wörterbuch*, in the right column the results of a search on the structural formulae.

The reference dictionaries for Middle High German are the *Mittelhochdeutsches Wörterbuch* (MWB) of the Göttingen and Mainz Academies of Sciences and Humanities, which is still being compiled and is also available as part of an online offering (MWB Online), and until the MWB is completed, its predecessors, the 19th-century dictionaries (BMZ, Lexer), which are accessible online in the Trier Dictionary Network. MWB Online also contains a complete lemma list of the precursor dictionaries and the MWB (approx. 90,000 words). The period for the sources of the older dictionaries extends into the 15th century, that of the MWB only up to 1350; it is estimated that the MWB will process around 50,400 words. The more comprehensive lemma list of the older dictionaries therefore provides a bridge to older Early New High German.

The *Altsächsisches Handwörterbuch* (ASWB) by Heinrich Tiefenbach (2010) and the *Mittelniederdeutsches Handwörterbuch* (MNWB) by Lasch/Borchling, which will shortly be completed, are not yet available online; they are to be digitized and published online in the Trier Dictionary Network as part of the project. The lemma list of the ASWB comprises around 6,900 words, that of the MNWB around 80,000 words.

1.5 Reference corpora

The reference corpora of the DeutschDiachronDigital initiative for Old German (ReA, OHG and OS), MHG (ReM) and MLG (ReN) cover approximately the same period as the reference dictionaries. ReM and ReN provide material structured according to time, space, and to some extent, text types. All three corpora are grammatically annotated (PoS, morphology) and lemmatized. They can be searched at the levels of tokens, annotations, and lemmas, but as with the dictionaries' online offerings, they are not interlinked. Linking with the reference dictionaries is performed only in the case of ReM (MWB, Lexer).

The vocabulary recorded in the reference corpora is (naturally) considerably smaller than that of the reference dictionaries: the ReA contains the complete textual tradition of OHG and OS but not the extensive tradition of glosses; it comprises 4,100 words for Old Saxon and 10,900 words for Old High German. The corresponding OHG and OS lemmas are not interlinked. ReM contains about 22,300 words, ReN about 17,000.

The linking of WoDia with the reference corpora enables the interconnection and joint use of the corpora, and it opens up an annotated textual evidence base for WoDia.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

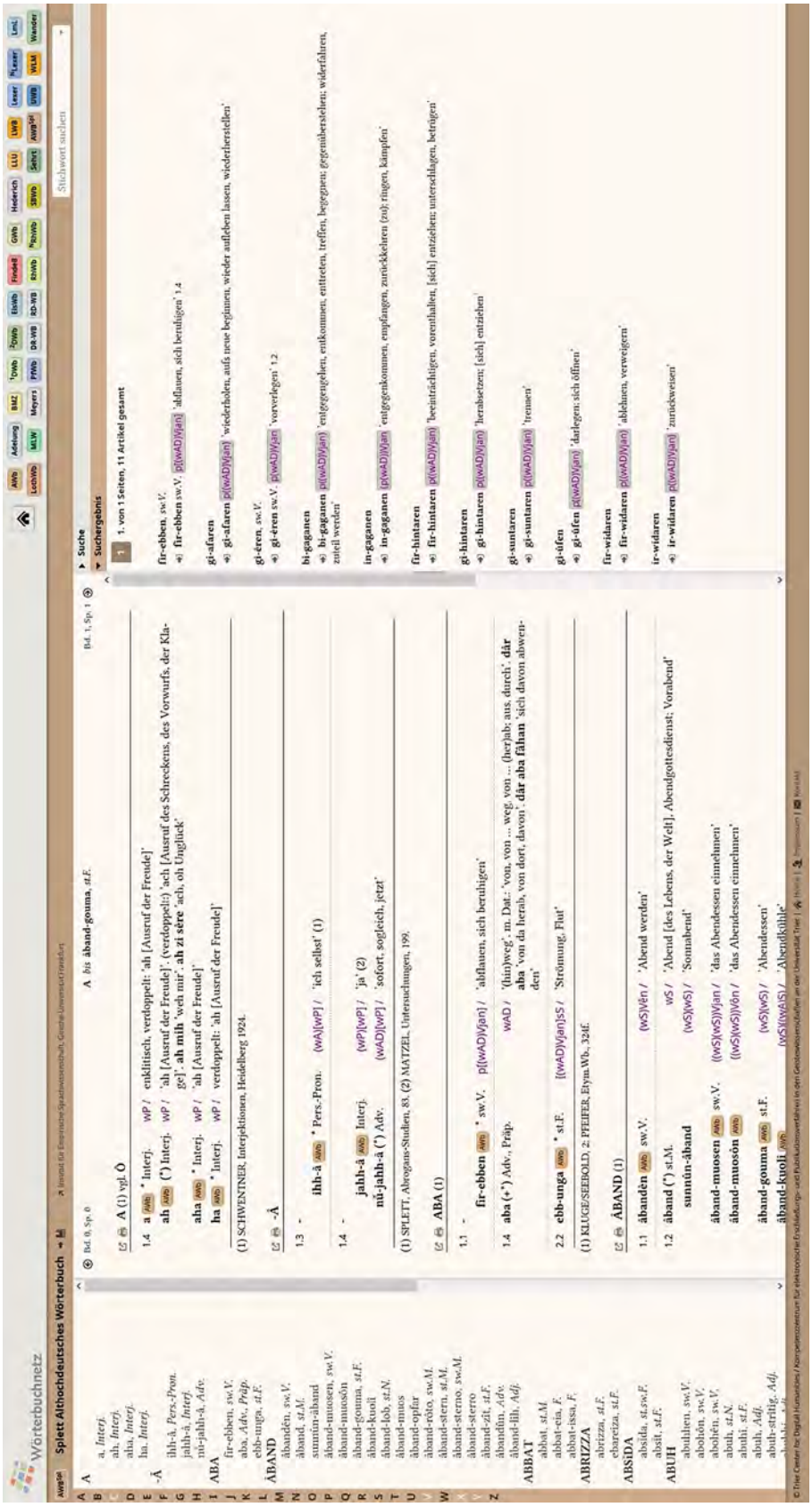

Fig. 1: Jochen Splett, *Althochdeutsches Wörterbuch* (1992), example page of the online publication (in preparation)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2. Basic data and work stages of the project

2.1 Basic ata

The project's basic data are the lemma lists of the reference dictionaries, for OHG additionally the word family structure and the structural formulae of AWB^{Spl}. Figure 2 summarizes the scope of the lemma lists included in WoDia (cf. 1.4 above) and the lemma lists of the reference corpora to be linked with them (cf. 1.5 above):

Variety	Dictionary	Reference Corpus
OS	6,900 (ASWB)	4,100 (ReA)
OHG	28,500 (AWB^{Spl})	10,900 (ReA)
MHG	90,000 (Lexer, to 15th century) including 50,400* (MWB to 1350)	22,300 (ReM)
MLG	80,000* (MNWB)	17,000 (ReN)

Fig. 2: Sizes (number of lemmas) and sources (dictionaries) of the vocabularies in WoDia; sizes of the vocabularies in the corresponding reference corpora; *= estimated number of lemmas when complete

2.2 Stages of work

In the following, the stages of work necessary for setting up WoDia are presented in a simplified version by comparison with the detailed project description of the DFG application.

2.2.1 Digitization of dictionaries and production of lemma lists

The two concise dictionaries ASWB and MNWB will be digitized and published in the Trier Dictionary Network. The lemma lists will then be extracted from the digitized data.

The lemma lists of OHG and MHG are already available. In the case of MHG, we shall mark that part of the vocabulary of the total lemma list from MWB + Lexer that is only attested after 1350 and thus falls within the scope of sources for the Early New High German dictionary that is currently being compiled. This is already apparent for the section so far processed in the MWB (i. e., lemmas of the complete list that, for reasons of dating, have not been included in the MWB); for the remaining parts, marking can be performed semi-automatically with the help of a list of the more recent sources in Lexer. Consequently, the vocabulary of older Early New High German is already incorporated in WoDia, and some groundwork for the inclusion of the Early New High German Dictionary's lemma list has already been performed. However, for reasons of scope and lack of preliminary work, this expansion of WoDia will have to be deferred for the time being.

2.2.2 Linking of the lemma lists

A prerequisite for the assignment to word families across linguistic epochs (cf. 2.2.5 below) is the linking of the lemmas that correspond etymologically. The linking of the OHG and MHG lemma lists has already been performed semi-automatically by Thomas Klein with the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

help of a suitable script (cf. Klein 2018, pp. 13–16). For the linking of the OS and MLG lists, similar scripts will be written.

2.2.3 Segmentation of the llemmas

For the morphological word-formation analysis, the lemmas are segmented into prefixes, stems, and suffixes on the basis of affix lists. OHG has already been analysed in AWB^{Spl}; for MHG, the analysis has been performed semi-automatically by means of a script by Klein 2018 (see above under 1.2). For OS and MLG, similar lists will be compiled and scripts developed.

2.2.4 Determination of word family stems

The stems obtained in 2.2.3 for OS and MLG are traced back to their WFSts by eliminating variants (umlaut, ablaut, grammatical change, gemination, etc.). For this, scripts similar to those used by Klein for MHG are needed. The rules for such elimination contained in these scripts must be developed anew for OS and MLG.

2.2.5 Assignment of word family stems to word families

Suggestions for the assignment to word families are generated for the WFSts determined in 2.2.4, using the direct correspondences already linked in 2.2.2 and with the help of the ZHistLex script for MHG (cf. above under 1.3) as well as the scripts that will be adapted for OS and MLG. OS WFSts will be linked to OHG word families and MLG WFSts to MHG word families.

Owing to the identical spelling of many WFSts, the suggestions are not only 1:1 correspondences but mostly multiple in nature so that these have to be examined individually; cf. for example Klein's (2018, p. 26) example of the homographic WFSt WINT, which belongs to the following word families: *winden* stV. (3a) 'to wind' – *wint* stM. 'wind' – *wint* stM. 'greyhound (Wendish dog)'.

If it is not possible to assign the OS and MHG WFSts to an OHG word family and the MLG WFSts to an MHG word family, equivalents are sought in the other varieties; if no word family can be assigned, the lemmas in question remain as individual words.

2.2.6 Analysis of the constituent structure / structural formulae

A central function of the research environment is the representation of the word families' inner structure, i. e., the stages of word formation that precede the composites and derivations, as expressed in Splett's work and elsewhere by means of bracketed formulas. As in Splett's work, the part of speech of the stems needs to be indicated in the formula. For a minority of cases in the vocabulary of MHG, OS, and MLG, where direct equivalents are present in OHG, the structural formulae of Splett can be transferred; for all other lemmas, the structural formulae have to be elaborated anew. This is done as far as possible through semi-automated processes, in which proposals are generated and then reviewed. A prerequisite for this is the segmentation of lemmas into stems and affixes (2.2.3) and the assignment of stems to word families (2.2.5). For example, a script for a complex lemma can search for freely occurring equivalents of its constituents within the same word family and adopt the specification for their part-of-speech or even a structural formula for the constituent that already exists through direct linking. Regularities that can be used for the semi-auto-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

matic analysis are already observed in the work on 2.2.5 and will be converted into corresponding scripts for the analysis of the constituent structure. In the processing, OS and MHG precede MLG, so that the already existing OS and MHG structural formulae can be transferred for the direct equivalents in MLG.

2.2.7 Setting up a web application

For the overall data management, a central database is created as a two-tier web application, with a component for data collection and modelling (back end) and one for publication and use (front end). The database is connected to the online dictionary resources and the reference corpora via open interfaces and web services. Basic functions of the front end are the display of word families (word family list, display of a specific word family, selection option for the varieties, etc.) and search functions for lemmas, word families, word formation structures, word formation elements, etc.

3. Conclusions

The establishment of WoDia marks significant progress in the development of a digital research infrastructure for German language history. This is done by bundling and supplementing existing digital resources (dictionaries), by introducing a new dimension of description and investigation (word families) across epochs and varieties, by the consistent use of semi-automatic procedures for the sophisticated analysis of extensive language data (vocabularies), and finally by making the results of the work available in a user-friendly web application for research and teaching, with links to the dictionary and corpus resources available online via web services.

References

ASWB = Tiefenbach, H. (2010): Altsächsisches Handwörterbuch. Berlin/New York.

AWB = Althochdeutsches Wörterbuch (1952 ff.): Auf Grund der von Elias v. Steinmeyer hinterlassenen Sammlungen im Auftrag der Sächsischen Akademie der Wissenschaften zu Leipzig. http://awb.saw-leipzig.de/cgi/WBNetz/wbgui_py?sigle=AWB (last access: 24-05-2022).

BMZ = Benecke, G. F./Müller, W./Zarncke, F. (1854–1866): Mittelhochdeutsches Wörterbuch. Leipzig. Digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/21. https://www.woerterbuchnetz.de/BMZ (last access: 24-05-2022).

DeutschDiachronDigital [Reference corpora on German language history]. https://www.deutschdiachrondigital.de/ (last access: 24-05-2022).

FWB = Frühneuhochdeutsches Wörterbuch. Bearbeitet von Oskar Reichmann et al. (1986 ff.): Berlin/New York. https://fwb-online.de/ (last access: 24-05-2022).

Klein, T. (2018): Mittelhochdeutsche Wortfamilien: Ermittlung und Perspektiven. In: Zeitschrift für Wortbildung/Journal of Word Formation 2/1, pp. 11–31. DOI: https://doi.org/10.3726/zwjw.2018.01.01 (last access: 24-05-2022).

Lexer = Mittelhochdeutsches Handwörterbuch von Matthias Lexer (18721878). Leipzig. Digitised version in the Dictionary Network of the Trier Centre for Digital Humanities, Version 01/21. https://www.woerterbuchnetz.de/Lexer (last access: 24-05-2022).

MNWB = Lasch, A./Borchling, C. (1956 ff.): Mittelniederdeutsches Handwörterbuch. Begr. von Agathe Lasch und Conrad Borchling. Kiel/Hamburg.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

MWB = Mittelhochdeutsches Wörterbuch (2006 ff.): Im Auftrag der Akademie der Wissenschaften und der Literatur Mainz und der Akademie der Wissenschaften zu Göttingen. Stuttgart. – MWB Online: http://www.mhdwb-online.de/index.html (last access: 24-05-2022).

Plate, R. (2020): Computergestützte Etablierung epochenübergreifender Wortfamilienstrukturen [ZHistLex-Teilprojekt]. Final report. https://zhistlex.de/ziele/wortfamilien/ (last access: 24-05-2022).

ReA (Reference corpus of Old Saxon and Old High German). https://www.deutschdiachrondigital.de/rea (last access: 24-05-2022)

ReM (Reference corpus of Middle High German): https://www.linguistics.rub.de/rem/ (last access: 24-05-2022)

ReN (Reference corpus of Middle Low German): https://www.slm.uni-hamburg.de/ren/ (last access: 24-05-2022)

Splett, J. (1993): Althochdeutsches Wörterbuch. Analyse der Wortfamilienstrukturen des Althochdeutschen. Zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes. 3 vols. Berlin/New York.

Splett, J. (2009): Deutsches Wortfamilienwörterbuch. Analyse der Wortfamilienstrukturen der deutschen Gegenwartssprache. Zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes. 18 vols. Berlin/New York.

Trier Dictionary Network: https://woerterbuchnetz.de (last access: 24-05-2022).

Contact Information

Ralf Plate
Akademie der Wissenschaften und der Literatur | Mainz, Mittelhochdeutsches Wörterbuch, Arbeitsstelle an der Universität Trier
plate@uni-trier.de

Acknowledgements

This article is based on the project description prepared jointly by Jost Gippert (Frankfurt), Sarah Ihden, and Ingrid Schröder (Hamburg), Thomas Burch (Trier), and the author. I should also like to thank Claudia Wich-Reif (Bonn) for her valuable comments. David Yeandle deserves my special thanks for translating the text from German into English.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.