Hauke Bartels

THE LONG ROAD TO A HISTORICAL DICTIONARY OF LOWER SORBIAN

Towards a lexical information system

Abstract The Sorbian Institute has been taking preparatory steps for a historical-documentary vocabulary information system for Lower Sorbian for about 10 years. To this end, the entire extant written material (16th–21st centuries) of this strongly endangered European minority language is to be systematically evaluated. An attempt made a few years ago to organise and finance the project as a long-term scientific project was not successful in the end. Therefore, it can only be advanced step by step and via some detours. The article informs about the interim status of the project, especially with respect to the creation of a reliable database.

Keywords Lower Sorbian; historical lexicography; minority language; e-lexicography; lexical information system; text corpus; Sorbian institute; language portal

1. Introduction

After 1945, there were two attempts to secure the necessary work programme for a comprehensive historical dictionary of Sorbian through an academic project: in the 1950s and 1960s by a working group around Hans Holm Bielfeldt at the Institute for Slavonic Studies of the German Academy of Sciences (Bielfeldt 1961; Müller 1967) and again in the 2010s, with a narrower focus on Lower Sorbian, by the Sorbian Institute (Bartels 2013).[1] Both attempts were ultimately unsuccessful.

Even if it is therefore not possible to fully implement the work programme last defined at the Sorbian Institute, it has not been completely abandoned. However, the circumstances now require a very long-term and small-scale structured approach, the outcome of which is quite open. Nevertheless, the project represents the first already advanced attempt to comprehensively record and, if possible, describe the historically transmitted vocabulary of Lower Sorbian on the largest possible data basis. The current article informs about the interim status of the project and gives an outlook on the next steps.

The overarching goal of all efforts is to document Lower Sorbian as comprehensively as possible, if feasible also in conjunction with language promotion measures. The language portal dolnoserbski.de serves this purpose. Some of the language resources gathered there are mentioned below. A visual impression is given by the following screenshot:

[1] The funding application was last submitted in 2016 via the Saxon Academy of Sciences and Humanities for the 2018 research programme of the Union of the German Academies of Sciences and Humanities under the title „SorbLex Lower Sorbian. A historical-documenting vocabulary information system of the Lower Sorbian language (internet »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.

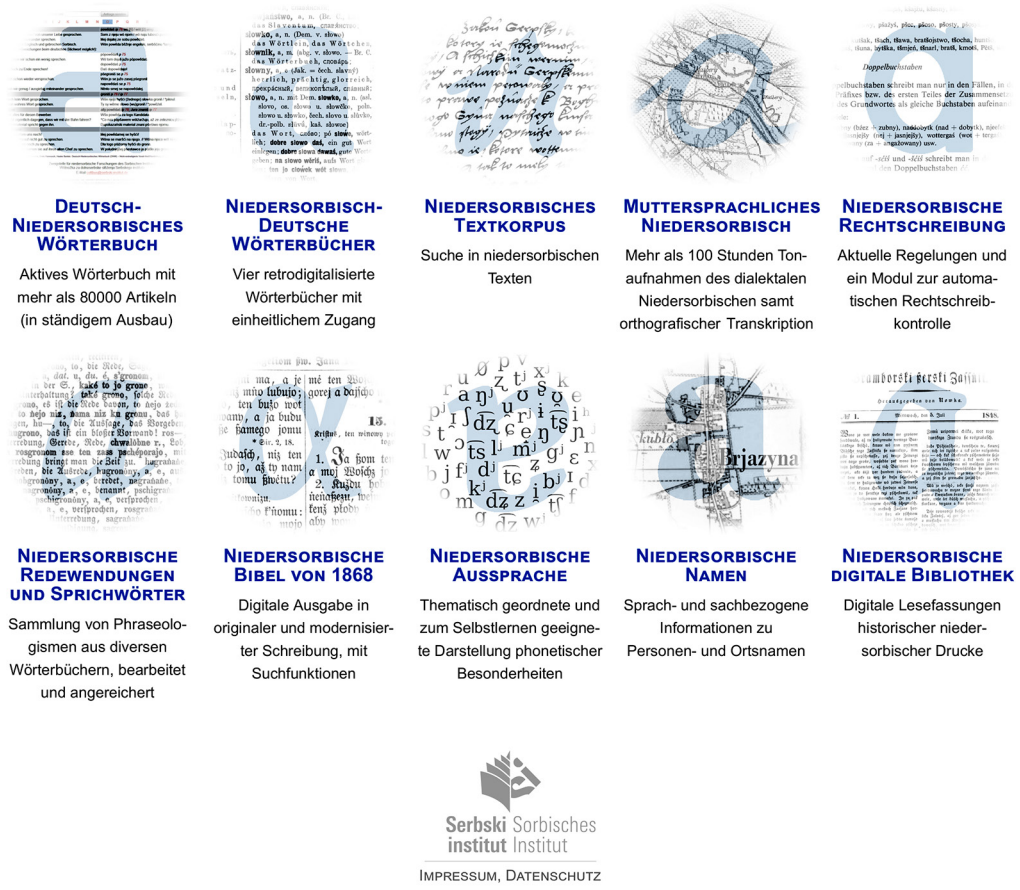

Fig. 1: The Lower Sorbian language portal (screenshot of the main page, German version)

2. Important stages on the road …

The work programme drafted for the second academy project was divided into three phases, which – among numerous others – envisaged three main tasks:

1) Preparation of the database
2) Inventory and rough description of previously unrecorded lexis
3) Systematic re-description of the entire historical vocabulary.

The third main task has been postponed indefinitely, and the number and scope of the tasks planned for phases 1) and 2) also had to be significantly reduced. Nevertheless, important progress has been made in recent years.

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.

2.1 The new Lower Sorbian text corpus

For Lower Sorbian, the aim is to build up a text corpus which – as a full historical corpus – comprises the entire extant printed literature from the first half of the 16th century to the present. The construction of such a corpus began in the mid-1990s. Until around 2015, the work had to be limited to the purely quantitative expansion of the corpus. Since then, the quality of the corpus texts has also been considerably improved in various projects, which is why we now speak of the "new" Lower Sorbian text corpus. (For the history and procedure, see Bartels 2020.) By 1945, the goal of a complete corpus had already been largely achieved. Individual gaps can probably be closed in the coming years. At the same time, the digitisation and processing of post-1945 literature is progressing so that a high-quality and almost comprehensive text corpus will soon be available as a database for historical lexicography.[2] The sub-corpus of literature from the mid-20th century onwards will contain the weekly newspaper "Nowy Casnik", which was quantitatively dominant in this phase, as well as a modest proportion of journalism and fiction, and a separate section of Lower Sorbian schoolbooks. The vocabulary expected in the last mentioned sources, including large amounts of terminology, is specific in several respects, but must definitely be included in the analysis because of its great influence on several generations of pupils and thus speakers. With a view to language change in general, but especially with a view to the socio-political upheaval that took place around and after German reunification in 1990 and which also had a strong linguistic impact in various ways, it is important that the corpus also covers the decades before and after this caesura as completely as possible.

Most of the collected and digitised written material up to 1945 is freely accessible via the Lower Sorbian language portal mentioned above. Since all units of the new text corpus are of high quality and have been structurally annotated (according to the TEI-P5 guidelines), they can directly serve as the basis for a digital library.[3] At the same time, the corpus texts are also directly available for queries via the so-called "convenience search".[4] This term emphasises the benefit of the previous preparation of the corpus texts for users who would like to research a historical corpus but do not have sufficient knowledge of the grammatical and orthographical variety of forms of the text words. In this way, the text corpus, which is at the same time an important source of knowledge for all kinds of Sorbian studies, should also be opened up to non-Sorabists with as few barriers as possible.

The preparation of the corpus text for this purpose, which currently ends with a (normalising) lemmatisation and still has to do without POS tagging, for example, is above all a prerequisite for the planned lexicographic evaluation. Expansion, analysis and annotation of the corpus will continue in order to create an optimal database. An important addition to this written language corpus is the approximately 100 hours of audio recordings made between 2011 and 2016 with native speakers of Lower Sorbian of mostly very high age. These recordings were fully transcribed and are therefore also easily accessible for lexicographic evaluation.[5]

[2] The entire text corpus currently comprises about 46 million tokens, the higher-quality "new" corpus about 25 million. The total size of an all-encompassing Lower Sorbian text corpus (16th century to 2020) can be estimated at 50 to 55 million tokens.

[3] See https://www.dolnoserbski.de/biblioteka/.

[4] See https://www.dolnoserbski.de/korpus/komfort/pytanje/.

[5] See https://www.dolnoserbski.de/dobes/.

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.

2.2 Dictionaries as part of the data base

The second data basis for a lexical information system of Lower Sorbian are digital and retro-digitised bilingual dictionaries.[6] A first step towards an information system (more on this term in the next chapter) was the retro-digitisation of the four most important Lower Sorbian-German dictionaries (Zwahr 1847; Mucke 1911–1928; Šwjela 1961; Starosta 1999). These were modelled in XML in a uniform, fine-granular manner and made accessible in 2012 via a common user interface.[7] They represent the essential part of 150 years of Lower Sorbian lexicographic tradition and their systematic inclusion ensures, among other things, the desired transparency with regard to previous descriptions of the vocabulary.

The usability of these sources was further increased in 2014 by indexing the German-language part (more than 580,000 word forms) in addition to the usual access via Lower Sorbian lemmas in Lower Sorbian-German dictionaries. Using this "German access", not only German translation equivalents can be found, but also words in descriptions of meaning, commentaries, etc. The German-language component of the dictionary text was largely orthographically normalised and lemmatised, so that a search for *Karussell* 'carousel', for example, also records the historical spelling *Carouffel*. Such a query immediately opens up a small panorama of Lower Sorbian expressions for this traditional fairground ride, such as *karusel* and the variant *karasel*, but also the alternative terms *ćwertaw(k)a*, *kóniki*, *hobćwertawka* as well as derivatives and phrases: *kónikowy*, *karaselać*, *na kónikach sejźeś*, *kóniki su pśišli*, *źomy na kóniki*. This approach is interesting not only for lexical but also for cultural-historical research, since Mucke's large dictionary in particular contains a great deal of "encyclopaedic" information.

Another component of these four retro-digitised dictionaries, which has been processed in a special way and thus made more accessible, are idioms and proverbs. These are numerous in the sources, but often difficult to find. Separate modelling and the addition of lexicographic information (for example, a literal German translation as well as explanations of meaning in Lower Sorbian and German), whose lexical components are also accessible via the search, make finding them much easier. Links lead directly from the references and the description to the sources.[8]

These four dictionaries as a lexicographic data source are supplemented by the new German-Lower Sorbian Internet dictionary (Deutsch-niedersorbisches Wörterbuch – DNW), which has been compiled for 20 years now and is constantly being expanded.[9] It currently contains more than 82,000 articles, some of them extensive, with numerous new Lower Sorbian expressions and a large number of example sentences (currently about 47,000). On the one hand, these, among others, make the DNW an active dictionary that thus makes an important contribution to language preservation. On the other hand, they can supplement the Lower Sorbian text corpus, admittedly as a very special part, since they were still predominantly formulated by native lexicographers.[10]

6 There are no monolingual dictionaries of Lower Sorbian.

7 See https://www.dolnoserbski.de/ndw/.

8 See https://www.dolnoserbski.de/nrs/.

9 See https://www.dolnoserbski.de/dnw/. See also Bartels (2010a).

10 The last native speaker lexicographer, Manfred Starosta, and another co-author of the DNW, Erwin Hannusch, have been retired for many years, but are still available as informants. Until 2008, 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.

2.3 Proper names

For Lower Sorbian proper names, a separate database has been built up in recent years, but it will be an integral part of the lexical information system. So far a good 1000 settlements or administrative units have been recorded. The site currently also contains information on more than 2600 surnames and about 120 forenames. Detailed information on the inventory, the sources evaluated, etc. can be found on the page itself.[11] Comprehensive information is provided on each of the proper names; selection and presentation vary according to the type of name. In the case of place names, in addition to more encyclopaedic information (including geo-coordinates and a short description on the affiliation to the Sorbian settlement area), there is information relevant to language comprehension (current, alternative and obsolete Lower Sorbian and German names and name forms) as well as information that promotes active use: derivatives (inhabitant, adjective) and the associated inflectional information in the form of a case-specific table. The entries for personal names contain the main form of the name (e. g. *Kóńcař*, *Měto* as a short form of *Mjertyn*), for family names additionally the name forms for women (*Kóńcarjowa*) and some other derivatives. In the case of first names, alternative and obsolete forms of the name, abbreviations and diminutives (in the above example, *Měćo*, *Mjećo*, *Mětko*) and augmentatives are also listed. Here too the adjectives formed by the respective name as well as specific inflection tables can be found. The site currently contains a total of more than 18,000 Lower Sorbian names and derived forms, most of them with complete inflection tables. For an effective search for names or name forms, more than 10,000 German names as equivalents of the Lower Sorbian ones are included as well. – There a plans to expand the proper names page in the future. A new function for conveying etymological information is currently being tested and further developed (Zschieschang 2021).

3. Lexical information system

In the academy application mentioned at the beginning, a detailed description was included of how a "historically-documenting vocabulary information system" of Lower Sorbian should be constructed conceptually and technically, what types of information it would contain, what forms of presentation it would offer, etc. The presentation of this concept for a polyfunctional and polyaccessive digital "dictionary", which most likely cannot be fully realised for the reasons mentioned, does not make sense here. Instead, we will briefly describe which aspects of a future information system already exist and which can probably be added to it in the coming years. Each of these additions brings the system a little closer to what the term "information system" promises.

The current Lower Sorbian language portal essentially comprises the individual resources already mentioned in chapter 2. On the one hand, these are part of the data basis for the development of a lexical information system, above all as a starting point for a new lexicographic description. On the other hand, they also represent a sub-component of the system, for example, as direct sources of previous lexicographic descriptions in the sense of lexicographic-historical transparency Bartels 2013,[12] p. 43).

work on the DNW was also supported by a circle of other mother-tongue informants. All but one of the members have since passed away.

[11] See https://www.dolnoserbski.de/mjenja/.

[12] The article very compactly presents an early version of the concept.

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.

So far, two aspects of the targeted information system are in the foreground of its gradual development: Firstly, the technical-conceptual linking of the individual resources. This should guarantee optimal availability of the information, independent of individual access routes. And secondly, the best possible solution of the so-called "finding problem" in historical lexicography (Reichmann 2012, p. 157 f.). In a similar way, however, this is also relevant for the use of historical corpora.

Both aspects are closely interwoven in many cases, which is why the steps taken so far often serve both goals at the same time:[13]

- the uniform data modelling and user interface for the Lower Sorbian-German dictionaries mentioned in chapter 2.2; in addition, the indexing of the German lexis contained there at various microstructure positions for "German access" to the dictionary information, and the additional information (and thus search options) for idioms and proverbs.
- a central Sorbian-language search on the portal's entry page, which initially includes the two most up-to-date dictionaries (Starosta 1999 and DNW).

Database and search mechanisms for the last mentioned step are designed in such a way that users can easily find the Lower Sorbian expression they are looking for (including multi-word expressions) even if they have only a very imprecise knowledge of the (historical) spelling. For this purpose, the system refers to graphically similar words. The query also recognises inflectional forms of Lower Sorbian words (such as those found in texts) and refers to their basic forms. Conversely, an additional search option also allows expressions with inflectional forms to be found when searching for the basic form. The query results appear in the form of word lists. For example, a search for *nan* ('father') produces a list of (currently) 29 single- and multi-word expressions, including such ones as *mama starego nana* ('great-grandmother'), which contain only an inflectional form of the search word. From these search results, one can then access the articles of the respective dictionaries, where further information on meaning, inflection, etc. can often be found.

Here it becomes apparent that for fully effective access to (historical) dictionary data as well as to the lexis from the historical full corpus, further data resources are necessary as part of the information system. For a strongly inflectional language such as Lower Sorbian, it is important to allocate the numerous inflectional forms to the respective basic form. For this purpose, an extensive database already exists with about 3.7 million inflectional forms for more than 90,000 basic forms (lexemes).[14] On the other hand, it is a matter of assigning the many historical spelling variants to a form representing this totality, for example in the sense of a construct lemma (Reichmann 2012, p. 160: "Konstruktlemma"). This database, which is currently being developed, will be the starting point for a corresponding reference system.

Further important steps in this direction would be the expansion of the central Sorbian search to the other Lower Sorbian-German dictionaries and a central German-language

[13] More information on these search options can be found on the portal pages themselves.

[14] At this point, it should be mentioned that Lower Sorbian, as a small or minority language, is also one of the so-called digitally under-resourced languages. This means that many resources cannot be adopted or purchased, but have to be developed specifically. This applies, for example, to an automatic spell checker, for which the database of Lower Sorbian inflectional forms was compiled, or – currently under development – a text-to-speech-function for Lower and Upper Sorbian.

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.

search in all these sources. Linking the (lemmatised) tokens from the text corpus (or from the texts of the digital library) with the dictionaries could also be another sensible step.

The originally planned complete lexicographical rewriting of Lower Sorbian vocabulary envisaged the creation of a comprehensive database that is uniformly described and modelled according to numerous other categories (for example, with regard to word formation components, word family affiliation, sense relations, etc.). On this basis, different variants of order or presentation should be made possible as well as an effective searchability of the entire dataset. It is still open how far this programme can be realised when we take the first steps towards a new description in the next few years.

4. Next steps

From what has been presented so far, it should have become clear that we already have a very good database due to the intensive work of the last decade. However, we will continue to work on supplementing and improving it. Some examples of ongoing and planned steps in this area have been mentioned above.

Beyond that, however, the lexicographic evaluation is to begin in 2023. This will initially involve identifying those words (including word formation and spelling variants as a basis for subsequent analysis) in the text corpus that have not yet been recorded lexicographically at all. This collection will be described for the first time. Among them there will be many loan words from German which, for purist reasons, have not been recorded in the dictionaries. The handling of these "non-Slavic" words in written Sorbian (Upper and Lower Sorbian), which in contrast to the dialects (vernacular) have often been replaced by words of Czech origin – mostly through the mediation of Upper Sorbian –, also plays a major role in the still unwritten history of Lower Sorbian.[15]

A longer-term goal is to trace individual stages in the development of the Lower Sorbian lexical system. To this end, it would be important, for example, to describe the lexical stock up to about the middle of the 19th century as it is reflected in the text corpus. Up to this time, the literature was almost exclusively religious (Bible, hymnbooks, sermons, etc.), whereas in 1848, with the first issue of the Lower Sorbian weekly newspaper "Bramborski Serbski Casnik"), other areas of lexis were also expanded or reflected in the literature. (Incidentally, even in this part of the literature, the way different authors deal with German loan words and other elements perceived as "vernacular" is already very different).

In addition to such an approach according to different periods of writing, a special treatment of culturally(-historically) significant parts of the vocabulary is also being considered (cf. e. g. Kämper 2016). It would be desirable if such a part of the upcoming lexical analysis could also be implemented. Despite all obstacles and detours, we are going our way. We just don't know yet how far we will get.

[15] There is no monograph on German loan words in Lower Sorbian as there is for Upper Sorbian (Bielfeldt 1933). Pohontsch (2002) is an important preliminary work for an overall presentation, because the replacement of German loanwords by Czech ones was in many cases carried out via Upper Sorbian. For loan words in Lower Sorbian, see also Bartels (2009, 2010b).

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.

References

Bartels, H. (2020): Das niedersorbische Globalkorpus als Ziel einer ganzheitlichen Konzeption zum Aufbau von Textkorpora. In: Lětopis 67 (2), pp. 3–44.

Bartels, H. (2013): Zur Konzeption eines historisch-dokumentierenden Wortschatz-Informationssystems des Niedersorbischen. Pläne zur Behebung eines drängenden Forschungsdesiderats. In: Kempgen, S./Wingender, M./Franz, N./Jakiša, M. (eds.): Deutsche Beiträge zum 15. Internationalen Slavistenkongress Minsk. München/Berlin/Washington, D.C., pp. 37–46.

Bartels, H. (2010a): The German-Lower Sorbian online dictionary. In: Dykstra, A./Schoonheim, T. (eds.): Proceedings of the XIV Euralex International Congress (Leeuwarden, 6–10 July 2010). Afûk, Ljouwert, pp. 1450–1462.

Bartels, H. (2010b): Das (diachrone) Textkorpus der niedersorbischen Schriftsprache als Grundlage für Sprachdokumentation und Sprachwandelforschung. In: Hansen, B./Grković-Major, J. (eds.): Diachronic Slavonic Syntax. Gradual Changes in Focus. München/Berlin/Wien, pp. 7–18.

Bartels, H. (2009): Loanwords in Lower Sorbian, a Slavic language of Germany. In: Haspelmath, M./Tadmor, U. (eds.): Loanwords in the world's Languages. A comparative handbook. Berlin/New York, pp. 304–329.

Bielfeldt, H. H. (1933): Die deutschen Lehnwörter im Obersorbischen. Leipzig.

Bielfeldt, H. H. (1961): Die sorabistischen Arbeiten des Instituts für Slawistik der Deutschen Akademie der Wissenschaften zu Berlin. In: Lětopis A, pp. 124–126.

Kämper, H. (2016): Kulturwissenschaftliche Orientierung in der Lexikologie. In: Jäger, L./Holly, W./Krapp, P. et al. (eds.): Language – culture – communication. An international handbook of linguistics as a cultural discipline. Berlin/Boston, pp. 737–747.

Mucke, E. (1911–15/1926, 1928): Wörterbuch der nieder-wendischen Sprache und ihrer Dialekte. Vol. 1: St. Petersburg 1911–15, Prag 1926; Vol. 2/3: Prag 1928.

Müller, K. (1967): Die Bedeutung des Sorbischen Thesaurus. In: Forschungen und Fortschritte. Berlin, pp. 283–285.

Pohontsch, A. (2002): Der Einfluss obersorbischer Lexik auf die niedersorbische Schriftsprache, Bautzen.

Reichmann, O. (2012): Historische Lexikographie. Ideen, Verwirklichungen, Reflexionen. An Beispielen des Deutschen, Niederländischen und Englischen. Berlin/Boston.

Starosta, M. (1999): Niedersorbisch-deutsches Wörterbuch. Bautzen.

Šwjela, B. (1961): Dolnoserbsko-němski słownik. Budyšyn.

Zschieschang, Chr. (2021): Onomastischer Wissenstransfer am Sorbischen Institut. Zwei neue Projekte. In Onomastica Lipsiensia 14, pp. 641–656.

Zwahr, J. G. (1847): Niederlausitz-wendisch-deutsches Handwörterbuch. (Photomechanical reprint. Bautzen 1989).

Contact information

Hauke Bartels
Sorbisches Institut
hauke.bartels@serbski-institut.de

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.

Acknowledgements

The progress described in this article on the "long road to a historical dictionary of Lower Sorbian" has only been possible over the years thanks to the dedicated and persistent cooperation of many people and the financial support of a whole series of projects by various institutions. My heartfelt thanks go to all staff and supporters. More information can be found on the pages of the Lower Sorbian language portal niedersorbisch.de.

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.