Hauke Bartels

Das niedersorbische GLOBALKORPUS als Ziel einer ganzheitlichen Konzeption zum Aufbau von Textkorpora

1. Einführung und Rahmen

Das Sorbische Institut (SI) arbeitet seit Mitte der 1990er-Jahre, dabei jedoch mit geringen finanziellen und personellen Ressourcen und daher nicht in der eigentlich gebotenen Intensität, am Aufbau elektronischer Textkorpora[1] (BARTELS 2010: 7 f., KAULFÜRST 2014a/b, WÖLKOWA 2014). Dieser Plural („Textkorpora“) wäre im Niedersorbischen ein Dual und steht für die beiden sorbischen Schriftsprachen Nieder- und Obersorbisch. Alle Bemühungen um den Korpusaufbau und -ausbau sind daher stets auf zwei zwar nah verwandte, aber dennoch selbstständige Sprachen gerichtet.[2]

Für das Sorbische (im Folgenden als Oberbegriff) und die sprachwissenschaftliche Sorabistik sind Textkorpora von herausragender strategischer Bedeutung. Anders als in großen und stabilen Sprachgemeinschaften sind sie nicht nur eine Datenquelle unter vielen, sondern – zumindest beim weit stärker bedrohten Niedersorbischen – für zahlreiche Fragestellungen bereits heute die einzige (zuverlässige) Informationsgrundlage. Sie stehen damit in direktem Zusammenhang mit Maßnahmen zur Sprachdokumentation (BARTELS 2012). Generelles Ziel ist daher die Zusammen- und Bereitstellung einer hochwertigen und vielfältig nutzbaren Datengrundlage für textkorpusbasierte Forschungen zum Sorbischen.

Wegen der begrenzten Kapazitäten liegt der derzeitige Tätigkeitsschwerpunkt nach wie vor auf Textdigitalisierung und -aufbereitung. Daher sind die Korpora, abgesehen von einzelnen Studien (z. B. BARTELS 2008) und der Konsultation bei lexikografischer Arbeit, noch eine eher potenzielle Forschungsgrundlage. Dieses Potenzial konnte bisher weder umfassend linguistisch noch darüber hinaus kulturwissenschaftlich (s. Kap. 2.2) genutzt werden. Eine Korpusaufbereitung, wie sie im Folgenden skizziert wird, soll eine breite Nutzung über dringliche sprachwissenschaftliche Aufgaben hinaus auch für nichtlinguistische sorabistische Fragestellungen ermöglichen und fördern.

Für Niedersorbisch sind die Arbeiten auf nahezu allen für den Korpusaufbau relevanten Handlungsfeldern weiter fortgeschritten, weshalb diese Sprache im Folgenden als Beispiel dienen wird. Der darzustellende organisatorische Gesamtrahmen sowie die nachfolgend beschriebenen Formate, Methoden und Vorgehensweisen gelten aber gleichfalls

---

[1] Mit Textkorpus ist hier eine digitale, im engeren Sinne in Form von Volltexten (daher „elektronisch“; vgl. Fußnote 4) vorliegende Menge von Texten als a u t h e n t i s c h e n – in diesem Fall schriftlichen – Sprachäußerungen gemeint. Das Textkorpus muss außerdem möglichst r e p r ä s e n t a t i v sein – in unserem Fall für die niedersorbische Schriftsprache; vgl. dazu Kap. 4.1) – sowie h i n r e i c h e n d u m f a n g r e i c h (vgl. Kap. 2.1 und 4.1). (STEFANOWITSCH 2020: 21 ff.).

[2] Zur Geschichte der Schriftsprachen s. z. B. FASSKE 1994; einen breiten aktuellen Überblick bieten MENZEL/POHONTSCH [im Druck].

für das Obersorbische. Der Artikel legt den Schwerpunkt auf grundlegende Fragen der Korpusaufbereitung für nicht ausschließlich sprachwissenschaftliche Zwecke unter den spezifischen Bedingungen der Sorabistik. Da der Text sich nicht ausschließlich an ein linguistisches Publikum richtet, werden auch Zusammenhänge dargestellt, die in mit Korpuslinguistik vertrauten Kreisen bekannt sind.[3]

Bis etwa 2014 stand der quantitative Ausbau des Textkorpus im Vordergrund. Dabei gelang es in den Jahren seit 2006 durch Drittmittelfinanzierung erhebliche Mengen von in Frakturschrift gedruckten niedersorbischen Texten vor allem aus dem 19. und frühen 20. Jahrhundert im Double-Keying-Verfahren und damit in hoher Volltext[4]-Qualität zu digitalisieren. Seit 2015 konnte dann, flankiert von mehreren Drittmittelprojekten (vgl. z. B. die Fußnoten 50 und 52), erstmals in größerem Umfang auch die qualitative Korpusaufbereitung angegangen werden. Wichtige Voraussetzung dafür war die in den Jahren zuvor erfolgte Erarbeitung umfangreicher lexikalischer Datenbestände durch die Retrodigitalisierung wichtiger Wörterbücher und die Verknüpfung dieser und weiterer Daten.[5] Die auf dieser Grundlage entstandene computerlesbare „lexikalische Datenbank" wird seitdem laufend erweitert und u. a. zur Bereitstellung einer Applikation zur automatischen Rechtschreibkontrolle genutzt (vgl. Kap. 4.4). Sie ist gleichzeitig eine wichtige Ressource für die Aufbereitung des Textkorpus und wird ihrerseits durch dieses Verfahren weiter angereichert. Insgesamt erfolgte in den letzten Jahren eine deutliche konzeptionelle Weiterentwicklung des Textkorpus. Das im Aufbau befindliche „neue" niedersorbische Textkorpus ist Gegenstand der folgenden Darstellung.

Die Verzahnung drittmittelgeförderter und grundständiger Projekte des Sorbischen Instituts ermöglichte erstmals den Einsatz einer (für SI-Verhältnisse) größeren Gruppe von Mitarbeiterinnen und Mitarbeitern für Korpusaufbau und -aufbereitung. Das längerfristige Vorhaben ist somit stark kollaborativ angelegt und mit Blick auf die vorliegenden Zwischenergebnisse eine kollektive Leistung.[6] Sofern einzelne Beteiligte für bestimmte Module des Gesamtprojekts besondere Verantwortung übernommen haben, wird dies in Fußnoten kenntlich gemacht.

---

[3] Der vorliegende Text geht zurück auf einen Vortrag des Autors auf der vom Institut für sächsische Geschichte und Volkskunde (ISGV), Dresden, im April 2018 veranstalteten Tagung zum Thema „Forschungsdesign 4.0 – Datengenerierung und Wissenstransfer in interdisziplinärer Perspektive". Der Vortragstitel lautete damals: „Von sprachwissenschaftlicher zu kulturwissenschaftlicher Nutzung: Der Aufbau eines interdisziplinär nutzbaren sorbischen Textkorpus". Andere Beiträge zur Tagung sind in einem Sammelband unter dem Tagungstitel veröffentlicht (KLINGNER/LÜHR 2019).

[4] Mit „Volltext" ist allgemein ein elektronischer, in einem Computersystem vorliegender Text gemeint. Beim Double-Keying-Verfahren wird eine Vorlage zunächst zweifach (durch zwei Personen) abgeschrieben. Die dadurch entstehenden Volltexte werden sodann computergestützt miteinander verglichen und die Abweichungen durch eine dritte Person überprüft. Auf diese Weise entstehen Textdigitalisate mit einer sehr geringen Fehlerquote (vgl. Fußnote 41).

[5] Es geht hier einerseits um das Deutsch-Niedersorbische Wörterbuch (DNW 2003–2020), andererseits um vier retrodigitalisierte niedersorbisch-deutsche Wörterbücher, deren Informationen feingranular im XML-Format modelliert (SZCZEPAŃSKI 2012) und über eine einheitliche Suchschnittstelle zugänglich gemacht wurden. Beide Ressourcen sind auf www.niedersorbisch.de zugänglich.

[6] In den letzten zehn Jahren waren gemeinsam mit dem Autor vor allem Fabian Kaulfürst und Marcin Szczepański auch konzeptionell am Gesamtvorhaben beteiligt. In den letzten Jahren – in unterschiedlicher Intensität und mit verschiedenen Aufgaben – zusätzlich auch Marek Slodička, Joanna Szczepańska und Thomas Menzel.

2. Perspektiven auf Textkorpora

Es ist naheliegend, dass beim Aufbau sorbischer Textkorpora zunächst mit denselben Fragestellungen und Herausforderungen zu rechnen ist wie bei vergleichbaren Vorhaben anderer, auch „großer" Sprachen. Andererseits gibt es, wie oben bereits angedeutet, sorabistische Spezifika.[7] Daher zunächst ein kurzer Exkurs zu verschiedenen Perspektiven auf Textkorpora.

2.1 Big Data vs. Small Data

Sammlungen maschinenlesbarer Texte entstanden zunächst als Grundlage für sprachwissenschaftliche Forschung. Einen wichtigen Entwicklungsschritt stellte Anfang der 1960er-Jahre das sog. Brown-Corpus[8] mit 500 Textausschnitten (Samples) und einem Gesamtumfang von etwa einer Million laufender Wortformen (Tokens[9]) dar. Seit Beginn der 1980er-Jahre markiert der Begriff „corpus linguistics" die Herausbildung einer eigenen Forschungsrichtung (Lüdeling/Kytö 2008, Stefanowitsch 2020). Vorbildcharakter für vergleichbare Korpora anderer Sprachen hatte das Anfang der 1990er-Jahre veröffentlichte British National Corpus (BNC) mit nun schon 100 Millionen Tokens.[10] Seitdem haben sich die verfügbaren Textkorpora stetig weiter vergrößert – so umfasst z. B. das Deutsche Referenzkorpus (DeReKo) am Institut für deutsche Sprache in Mannheim mittlerweile 46,9 Milliarden „Wörter" (Stand 18.01.2020; Internet: https://www1.ids-mannheim.de/kl/projekte/korpora [20.04.2020]).

Mit dieser Tendenz zu immer größeren Textkorpora, die dennoch bei großen Sprachen stets nur einen kleinen Teil des Schrifttums repräsentieren können, ging – nicht zwingend, aber logisch wie pragmatisch nachvollziehbar – eine Tendenz zu „Big Data"-Ansätzen bei Aufbereitung und Auswertung einher. Sehr vereinfacht gesprochen, dabei grob typisierend und stark zugespitzt, stehen sich (u. a. im Hinblick auf Verfahren zur Korpusaufbereitung) verschiedene „Denktraditionen" gegenüber, deren (gelegentliches) Gegeneinander bzw. (wohl häufiger) Nebeneinander nicht selten zu Kontroversen oder gar Konflikten über das „richtige" Vorgehen führt. Vielleicht ist es zulässig, diese zwei Pole mit „automatische Sprachverarbeitung" einerseits und „Philologie" andererseits zu benennen. Ihnen lassen sich jeweils bestimmte Schlagworte, die z. B. für bestimmte methodische Traditionen stehen, sowie Einstellungen zuschreiben (ohne Anspruch auf Vollständigkeit):

---

[7] Diese sind aber nicht nur „sorabistisch", sondern vielmehr typisch für Kleinsprachen, die zudem häufig „digital unterversorgt" sind, d. h. nicht über ein hinreichend breites Spektrum an sprach- und korpustechnologischen Ressourcen verfügen.

[8] Brown University Standard Corpus of Present-Day American English. Internet: https://en.wikipedia.org/wiki/Brown_Corpus [15.07.2020].

[9] Als Token gilt in der Korpuslinguistik die kleinste zählbare Einheit, wobei die Frage, was konkret als Token zu betrachten ist, für ein Korpus jeweils definitorisch festgelegt wird. In unserem Fall gilt – grob gesprochen – ein laufendes, i. d. R. durch Leerzeichen isoliertes Textwort als Token (vgl. Kap. 4.4.2).

[10] Siehe die Website des BNC: www.natcorp.ox.ac.uk [15.07.2020].

Automatische Sprachverarbeitung	– „Big Data"/Quantität/distant reading[11] – möglichst weitgehende Automatisierung – relativ große Fehlerakzeptanz mit Blick auf einzelne Textstellen[12]
Philologie	– Fokus auf Einzeltexten/Qualität/close reading – Skepsis gegenüber Automatisierung – keine/geringe Akzeptanz für „Datenrauschen"

Beide Herangehensweisen kommen selten in Reinkultur vor. Außerdem haben beide Ansätze zweifelsohne auf bestimmten Handlungsfeldern ihre volle Berechtigung. Methodische Kombinationen und Mischformen sind zulässig und in bestimmten Situationen geboten. Gerade mit Blick auf die immer größer werdenden verfügbaren Datenmengen und darauf gestützte „datengeleitete strukturentdeckende Verfahren" (SCHARLOTH 2018: 66) ist eine „philologische" Textaufbereitung tatsächlich nicht realistisch:

> Die Datenmengen, mit denen nach unseren Vorstellungen im Rahmen datengeleiteter Analysen gearbeitet werden sollte, sind viel zu umfangreich, als dass sie noch durch Lektüre erschlossen, geschweige denn aufwendig kodiert werden könnten. […] Wir sind überzeugt, dass an Arbeiten auf dem Gebiet der Gesellschaftsanalyse qua Sprachanalyse künftig immer mehr die Forderung nach einer breiten empirischen Basis (‚big data') herangetragen wird, einer empirischen Basis, die sich einer qualitativen Bearbeitung von vornherein verschließt. (SCHARLOTH/EUGSTER/BUBENHOFER 2013: 349, 377)

Und sie ist mit Blick auf bestimmte Verfahren, vor allem „makroskopische" (GRAHAM/MILLIGAN/WEINGART 2016) statistische Auswertungen, wohl auch nicht notwendig: „Statistische Analyseverfahren erweisen sich gegenüber verrauschten Daten, die etwa durch automatische Texterkennungsverfahren entstehen können, mitunter als recht fehlertolerant." (BUBENHOFER/ROTHENHÄUSLER 2016: 67) Und auch für eine korpusbasierte Lexikografie großer Sprachen wird es mehr und mehr zur Herausforderung, die übergroße Datenmenge beherrschbar zu machen: „Data-sparseness is becoming a thing of the past; if anything, the challenge now is to devise means of preventing lexicographers from drowning in too much information" (RUNDELL/ATKINS 2013: 1336).

Andererseits hängen die Angemessenheit des gewählten Verfahrens und die Anforderung an Textkorpora von den Rahmenbedingungen und Anwendungszwecken ab. Sprachhistorische Forschungen etwa erfordern zusätzliche Maßnahmen: So konstatiert Alexander Lasch mit Blick auf deutsche Korpora:

---

[11] Mit den Begriffen „close reading" vs. „distant reading" sind zwei verschiedene Handlungsweisen bei der Deutung von Texten gemeint: erstere entspricht dem traditionellen (z. B. literaturwissenschaftlichen) Herangehen mittels genauer Lektüre, Analyse und Interpretation vor dem Hintergrund der Struktur des gesamten Textes. Beim „distant reading" wird von der Textstruktur abstrahiert und auf einzelne Elemente fokussiert, die dann auch über Textgrenzen hinweg, d. h. zahlreiche Texte auf einmal analysierend, in den Blick genommen werden (JÄNICKE et al. 2015).

[12] „From experience […] it can also be said that success rates beyond 70 % are usually good enough to provide an acceptable level of text search through a presentation system." (Zitiert nach MASTERPLAN ZEITUNGSDIGITALISIERUNG 2017: S. 34.)

> [Zwecks sprachhistorischer Forschung – Ergänzung H. B.] sind für die Erschließung ‚virtueller Korpora'[13] durch maschinelle Analyse Annotationen von hoher Qualität notwendig, die eine standardisierte Übertragung ins Neuhochdeutsche mit einschließen. Ohne diese Standardisierung sind Entwicklungslinien von Sprache anhand sprachlicher Muster, Konstruktionen, kaum plausibel nachzuvollziehen. Korpora als Datensammlungen, die diesem Anspruch genügen, gibt es bis zum jetzigen Zeitpunkt nicht; sie wären die Grundlage ‚virtueller Korpora', die in Bezug auf ein bestimmtes Thema zusammengestellt werden. (Lasch 2014: 236)

Für das Sorbische haben wir es außerdem nicht mit einer ansonsten problematischen Informationsflut (information overload, vgl. Tarp 2012: 255) zu tun, der wir irgendwie Herr werden müssten. Ganz im Gegenteil kann eine zumindest mit Blick auf bestimmte Zielstellungen und Auswertungsmethoden wegen des geringen Umfangs beschränkte Nutzbarkeit[14] von Kleinsprachen-Korpora nur durch eine optimale Erschließung teilweise ausgeglichen werden. Zumindest beim Niedersorbischen besteht das Problem daher nicht in einem Zuviel, sondern eher in einem Zuwenig an Daten.

Angesichts der daraus resultierenden Relevanz a l l e r überlieferten Daten für die sorabistische Forschung und eines relativ kleinen – und daher mit Blick auf den Aufbereitungsaufwand beherrschbaren – Umfangs des digitalisierten sorbischen Schrifttums soll mit dem hier dargestellten Vorgehen eine möglichst ausbalancierte Verbindung von Standards und Verfahren der automatischen Sprachverarbeitung mit einem größtmöglichen Maß an philologischer Original- und Detailtreue versucht werden.

2.2 Von sprach- zu (auch) kulturwissenschaftlicher Korpusnutzung

Zu einer Weitung des Blicks auf Textkorpora trug und trägt zudem die Herausbildung der Digital Humanities (Schreibman/Siemens/Unsworth 2016; Berry/Fagerjord 2017) bei. Impulse aus Computer- und Korpuslinguistik nutzend, führt dieser Prozess zu einer Erweiterung des Methodenrepertoires der (verstärkt digitalen) Geisteswissenschaften und lenkt auch jenseits der Sprachwissenschaft deren Interesse auf Textkorpora als möglicherweise vielversprechende Datenquelle. Dabei gelten insbesondere Korpus- und Diskurslinguistik als „the twin pillars of language research" (Sinclair 2004: 11; zitiert nach Taylor/Marchi 2018: 1). Beispiele für nicht-linguistische Fragestellungen an sorbische Textkorpora wären etwa: Ab wann taucht der Begriff „Minderheit" im Schrifttum auf, wann entsteht ggf. ein „Minderheitendiskurs"? Wie gestaltet und verändert sich dieser? Ab wann und wie wird die „Sprachfrage" (Bedrohung, Purismus usw.) explizit

---

[13] Damit ist eine thematisch zusammengestellte Teilmenge eines Basis-Textkorpus als Repräsentanz eines Diskurses gemeint.

[14] Dies gilt für viele Bereiche, so zum Beispiel schon für die nicht nur lexikografisch relevante Ermittlung von Mehrwortausdrücken. In einer Darstellung etablierter computergestützter Verfahren der Korpusauswertung zur Ermittlung von Kandidaten für lexikalische Mehrwortausdrücke in Form einer sog. „acquisition pipeline" betont Evert (2013: 1417): „It should be emphasised at this point that the word combinations identified by an acquisition pipeline can be meaningful and complete only to the extent that the base corpus from which they were extracted is representative of the underlying language, o f s u f f i c i e n t s i z e , a n d a n n o t a t e d w i t h r e l e v a n t l i n g u i s t i c i n f o r m a t i o n ." [Hervorhebung H. B.]

thematisiert? Wie werden Personen, Orte und Ereignisse dargestellt? Wie spiegeln sich historische politische Umbrüche wider, und gehen sie mit „sprachlichen Umbrüchen" einher?[15] Wie ist das Verhältnis identifizierter sorbischsprachiger zu (zeitlich oder inhaltlich parallelen) deutschsprachigen Diskursen?

Jedenfalls verspricht die Verbindung[16] von Korpuslinguistik, Diskurslinguistik und den sich Textkorpora gegenüber öffnenden Kulturwissenschaften[17] neue Zugänge und Erkenntnisse – sofern die jeweils verfügbare Textgrundlage auf eine Weise aufbereitet ist, die Derartiges[18] ermöglicht. Die skizzierte Entwicklung erhöht somit die Bedeutung von Textkorpora für eine multidisziplinär aufgestellte „Digitale Sorabistik" als Teil der Digital Humanities weiter.[19] Dies gilt umso mehr, als es mit Blick auf ein relativ überschaubares überliefertes Schrifttum im Unterschied zu anderen Sprachen sogar möglich ist, ein historisches Vollkorpus (s. Kap. 4.1) zu schaffen, also quasi auf (nicht ganz einfach zu gewährleistende) Repräsentativität zu verzichten zugunsten einer vollständigen Erfassung der Grundgesamtheit. Diesem Ziel sind wir beim Niedersorbischen bereits sehr nahe. Somit steht bald – bei rechtlichen Beschränkungen zumindest intern – nahezu das gesamte überlieferte niedersorbische Schrifttum als einzigartige Datenbasis für sprach- wie kulturwissenschaftliche korpusanalytische Untersuchungen zur Verfügung.

3. Sorbische Textkorpora: Zielsetzung und Entwicklungsrahmen

Vor dem Hintergrund des bisher Gesagten und mit Blick auf die benannten Spezifika

1. einer herausragenden Bedeutung digitalisierter Texte für die sorabistische Forschung,
2. der relativen quantitativen Begrenztheit des (digitalisierten) sorbischen Schrifttums,
3. der zunehmenden Bedeutung von Textkorpora auch für nicht-linguistische Forschung,
4. der sehr begrenzten finanziellen wie personellen Ressourcen für die Korpuserstellung und infolgedessen auch
5. der bisher nur in unbefriedigendem Maße erfolgten Korpusauswertung

gelten für den Korpusaufbau folgende Handlungsmaximen:

---

[15] Siehe für entsprechend ausgerichtete Untersuchungen z. B. KÄMPER 2007 sowie allgemein zu diskurslinguistisch repräsentierten Instanzen des „kollektiven Gedächtnisses" als „Gegenstand einer integrierten Kulturanalyse" KÄMPER 2015.

[16] Eine Verbindung, die letztlich in einem Aufgehen münden könnte: „Am Ende einer beeindruckenden Erfolgsgeschichte ist die Korpuslinguistik somit dabei, sich aufzulösen und in die Digital Humanities-Bewegung zu integrieren." (MAIR 2018: 24)

[17] Gleichzeitig haben sich Teile der Sprachwissenschaft und auch die Korpuslinguistik in den letzten Jahrzehnten zunehmend kulturwissenschaftlich bzw. „kulturanalytisch" ausgerichtet (BUBENHOFER/SCHARLOTH 2016, SCHRÖTER et al. 2019).

[18] „Korpuspragmatisch zu forschen bedeutet [...], in großen Textkorpora induktiv nach signifikant häufig auftretenden Mustern zu suchen und diese Muster als Ausdruck von rekurrenten Sprachhandlungen [...] zu interpretieren, mithin als Muster mit soziokultureller Salienz." (SCHARLOTH 2018: 65)

[19] Die Konzeption einer „Digitalen Sorabistik" soll in einem späteren Artikel ausgelotet werden.

1. Das sorbische Schrifttum soll möglichst vollständig und in höchstmöglicher Qualität digitalisiert und verfügbar gemacht werden.
2. Bei hohen Qualitätsansprüchen soll die Aufbereitung mit größtmöglicher Effizienz und Nachhaltigkeit erfolgen.
3. Die Korpusaufbereitung und -bereitstellung soll eine effektive wie auch breite Nutzung dieser Datenressource ermöglichen und damit die korpusbasierte Forschung zum Sorbischen (im weitesten Sinne) fördern.
4. Vor allem mit Blick auf nicht-linguistische Nutzungen sollen Zugangsbarrieren[20] zu den Daten möglichst abgebaut werden.

Bestimmte Nutzungsbarrieren sind mit vertretbarem Aufwand (derzeit) nicht abzubauen: So wird man bis auf weiteres für einen qualifizierten Zugang zu sorbischen Textkorpora ein bestimmtes Maß an Sprachkenntnissen voraussetzen müssen. Allerdings wird jenseits des traditionellen linguistisch-sorabistischen Nutzungskontextes schnell klar, wie stark die bisherige, noch nicht sehr weit gehende (vgl. Kap. 4.4) Korpusaufbereitung einer breiteren Nutzung entgegensteht. Dies wird deutlich, wenn man sich vergegenwärtigt, dass für korpuslinguistische Auswertungen Finden und Zählen grundlegend sind. Zentrale Fragen für die auf Vorkommenshäufigkeit basierenden und damit frequenzorientierten etablierten Methoden sind solche wie:

– Wann und wie häufig kommen bestimmte „Ausdrücke“ oder Kombinationen von Ausdrücken (Kookkurrenzen) vor?
– Wo ist ein überzufälliges und rekurrentes Auftreten von Wortverbindungen festzustellen? (Zum Beispiel als mögliche Identifikation von „Sprachgebrauchsmustern“ und damit von „Kristallisationskernen von Diskursen“, Bubenhofer 2009.)

Wie man einen zu findenden, zu zählenden, zu deutenden usw. „Ausdruck“ definiert, wird damit zum Schlüssel für ein Textkorpus, und zwar unabhängig davon, ob man aktiv mit Hilfe eines Korpus-Analyse-Programms einen konkreten Ausdruck sucht (Kommt „Stalin“ vor? Wo und wie oft? usw.) oder ob man einen eigens programmierten Algorithmus nach Kombinationen von Ausdrücken oder Mustern suchen lässt – denn auch dafür muss definiert werden, was gesucht und ggf. wie gezählt werden soll. Und dies kann dann auch die Grundlage sein für weiterführende Analysen großer Datenmengen sowie für Visualisierungen als „eigenständige Mittel der Erkenntnisgewinnung“ (Bubenhofer/Scharloth 2016: 930; s. auch Bubenhofer/Kupietz 2018). Was ist also mit Blick auf ein breit nutzbares niedersorbisches Textkorpus, das perspektivisch die überlieferte historische Tiefe möglichst komplett (als historisches Vollkorpus, s. Kap. 4.1) abdecken soll, an zusätzlicher Annotation notwendig, um vielen Interessierten mit den notwendigen Mindestsprachkenntnissen zu ermöglichen, weitgehend zuverlässig finden[21] und zählen zu können?

---

[20] Dieses Problem thematisieren in anderem Zusammenhang auch Blessing et al. (2015: 3f.). Dabei geht es auch um die Frage, wie durch möglichst zuverlässige Texte und Zugriffsinfrastrukturen vermieden wird, dass sich „die Aussicht auf leichten Zugang zu großen Textmengen oft als Falle“ erweist.

[21] Auch aus der Lexikografie ist dieses „Findeproblem“ bei der Nutzung historischer Wörterbücher bekannt, da mit einer in Bezug auf Schreibvarianten usw. sehr unterschiedlich ausgebildeten „Identifikationskompetenz“ gerechnet werden muss (Reichmann 2012: 161 f.).

3.1 Begriffsdefinitionen

Die Darstellung der vielfältigen Aspekte eines längerfristigen und kooperativen Vorhabens zum Korpusaufbau und -ausbau erfordert terminologische Differenzierungen in einem von begrifflichen Mehrdeutigkeiten geprägten Umfeld. Im Folgetext werden daher einige Begriffe terminologisch verwendet und durch Kapitälchen markiert. Es handelt sich um folgende Termini (vgl. Kap. 4.3 und 4.4):

- ROHTEXT: Digitaler Volltext unterschiedlicher Herkunft und Beschaffenheit, der noch nicht hinsichtlich seiner Korpuseignung geprüft und noch nicht bis zur KORPUSREIFE bearbeitet wurde.
- KORPUSREIFE: Definierte(r) Bearbeitungsstand und Qualitätsstufe eines ROHTEXTES; markiert die Schwelle zum
- KORPUSTEXT: ROHTEXT, der mit Blick auf seine beabsichtigte Aufnahme in das GESAMTKORPUS mehrere Prüf- und Aufbereitungsschritte (s. Kap. 4.3) durchlaufen und so KORPUSREIFE erlangt hat.
- GESAMTKORPUS: Gesamtmenge aller KORPUSTEXTE
- FORSCHUNGSKORPUS: Teilmenge des GESAMTKORPUS, die mit Blick auf einen bestimmten Verwendungszweck zusammengestellt und ggf. gesondert aufbereitet (z. B. annotiert) wurde.
- GLOBALKORPUS: Herausgehobene Form eines FORSCHUNGSKORPUS, das mit Blick auf eine möglichst breite und transdisziplinäre Nutzung bestimmte allgemeine Aufbereitungs- und Annotationsschritte durchlaufen hat.

Dagegen werden hier nicht explizit definierte Begriffe wie *Volltext*, *Korpus*, *Textkorpus* sowie diverse darauf basierende Wortbildungen weiterhin eher informell im Rahmen ihrer üblichen Verwendungsbreite gebraucht. Insbesondere der Begriff *Textkorpus* wurde nicht terminologisch verengt, da seine häufige und unbestimmte bzw. breite Verwendung auch im Sinne von *Textsammlung* kaum vermieden werden kann. So ist auch die Rede vom „neuen" vs. „alten" niedersorbischen Textkorpus möglich, obwohl das „alte" Textkorpus keine KORPUSTEXTE im terminologischen Sinne enthielt – die Unterscheidung „altes/neues Textkorpus" wird im Folgenden klar werden.[22]

Diese terminologische Lösung erhebt keinerlei Gültigkeitsanspruch über den vorliegenden Text bzw. die Diskussionen in der am Korpusaufbau beteiligten Arbeitsgruppe am Sorbischen Institut (vgl. Fußnote 6) hinaus. Die gewählten Termini mögen außerdem diskutabel sein, sie schaffen aber wenigstens intern (größere) Klarheit. Da der Zweck eines FORSCHUNGSKORPUS, das umfassend, vielseitig, transdisziplinär und dennoch effektiv nutzbar sein soll, ein besonderer ist, schien auch eine gesonderte Benennung nützlich und die Wahl fiel auf GLOBALKORPUS – auf Grundlage der (laut Duden) Bedeutung von *Global-* als ‚umfassend; nicht ins Detail gehend, allgemein', wobei mit Blick auf die Funktion und Aufbereitung der enthaltenen Texte die zweite Bedeutungskomponente (‚nicht ins Detail gehend, allgemein') entscheidend ist.

---

[22] Für eine schnelle Kurzinformation s. Kap. 5.2.1 und 5.2.2.

3.2 Entwicklungsrahmen für sorbische Textkorpora: GESAMTKORPUS – FORSCHUNGSKORPUS – GLOBALKORPUS

Die bisherigen sorbischen Textkorpora, die im Sinne der hier vorgestellten Neukonzeption zunächst „digitale Textsammlungen" waren und teilweise noch sind, wurden über Jahre weitgehend unabhängig voneinander und teilweise auf sehr unterschiedliche Weise aufgebaut (s. die Literaturhinweise in Kap. 1), sodass eine wichtige Aufgabe die nachträgliche Vereinheitlichung aller enthaltener Texte ist. So gilt der Entwicklungsrahmen für beide sorbische Sprachen, auch wenn das Niedersorbische in diesem Artikel im Vordergrund steht. Alle KORPUSTEXTE – niedersorbische wie obersorbische – gehen in das sorbische GESAMTKORPUS ein. Dieses bildet einen gemeinsamen konzeptionellen und organisatorischen Rahmen für die weitere Korpusaufbereitung und -entwicklung:

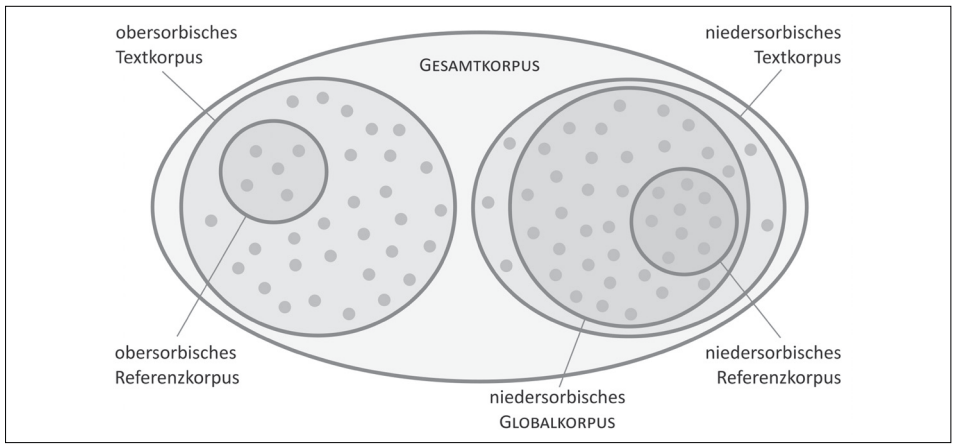


Abb. 1: Überblick sorbische Textkorpora

Alle KORPUSTEXTE im GESAMTKORPUS sind eindeutig einer der beiden Schriftsprachen zugeordnet (vgl. 4.4.6). Die Konzepte GESAMTKORPUS und obersorbisches bzw. niedersorbisches Textkorpus unterscheiden sich somit nur hinsichtlich ihrer Sprachzuordnung, ansonsten bilden sie die Gesamtheit aller KORPUSTEXTE, deren Herstellung und Weiterbearbeitung im Folgenden dargestellt wird.

Für das Niedersorbische existieren derzeit zwei FORSCHUNGSKORPORA: ein niedersorbisches Referenzkorpus, das als erstes Abbild des zeitgenössischen niedersorbischen Schrifttums (1990–2010) dient,[23] sowie der Grundstock[24] für das angestrebte

[23] Das gleiche gilt für das obersorbische Referenzkorpus, vgl. Kap. 4.6.

[24] Damit sei zugleich klargestellt, dass die Kreisgrößen in Abb. 1 nicht die tatsächlichen derzeitigen Mengenverhältnisse darstellen, sondern eher Zielmarken der Korpusaufbereitung. Auch die Größenrelation zwischen den beiden sorbischen Textkorpora entspricht nicht der Realität: Derzeit ist das intern zugängliche niedersorbische Textkorpus – vgl. Kap. 5.2 für den öffentlichen Teil – mit etwa 43 Millionen Tokens etwas größer als das obersorbische. Längerfristig sollte aber das obersorbische Textkorpus erheblich umfangreicher sein als das niedersorbische (vgl. Fußnote 31).

niedersorbische[25] GLOBALKORPUS. Dies wird im Mittelpunkt der folgenden Darstellung stehen.

4. Aufbau des niedersorbischen Textkorpus

In diesem Kapitel wird zunächst ein Überblick über den Gesamtprozess des Korpusaufbaus gegeben, wie er sich aktuell darstellt. Anschließend wird ab Abschnitt 4.4 die schon im Titel formulierte weitergehende Aufbereitung der KORPUSTEXTE für ein niedersorbisches GLOBALKORPUS thematisiert.

4.1 Das Textkorpus als Repräsentant des niedersorbischen Schrifttums

Wie bereits erwähnt, gilt für Niedersorbisch das erklärte Ziel, ein Textkorpus aufzubauen, das – als historisches Vollkorpus – das gesamte erhaltene[26] gedruckte Schrifttum von der ersten Hälfte des 16. Jahrhunderts bis in die Gegenwart annähernd vollständig[27] umfasst. Dieses Ziel wird zurzeit prioritär für alle bis 1945 erschienenen Druckschriften verfolgt, mit Blick auf publizistische Texte auch schon darüber hinaus.

Bei diesem Ansatz beantworten sich zentrale Fragen des Korpusdesigns (HUNSTON 2008, insbesondere S. 160–162), die vor allem Repräsentativität, Ausgewogenheit und Größe betreffen – zumindest mit Blick auf das gesamte niedersorbische Textkorpus und von erklärten Abweichungen abgesehen – von selbst: Die Korpusgröße entspricht dem Umfang aller Textdigitalisate. Die (Un-)Ausgewogenheit des Korpus entspricht der (Un-)Ausgewogenheit des Schrifttums (vgl. Abb. 2). Und das Korpus stellt keine wie auch immer definierte Stichprobe dar, sondern ein Abbild der Grundgesamtheit. Bezüglich der auf dieser Basis erstellten FORSCHUNGSKORPORA sind diese Fragen selbstverständlich mit Blick auf die jeweils verfolgten Forschungsziele zu beachten.

Die notwendige Grundlage für eine in diesem Sinne begründete Zusammenstellung des Vollkorpus ist eine möglichst umfassende Inventarisierung des bekannten Schrifttums.[28] Diese Aufgabe wurde in den letzten Jahren, zunächst für den Zeitraum bis 1945,

---

[25] Für das Obersorbische ist die Erarbeitung eines GLOBALKORPUS v. a. wegen der gänzlich anderen Situation im Bereich der Primärdigitalisierung und damit der Qualität verfügbarer ROHTEXTE derzeit noch nicht möglich.

[26] Historisch kam es mehrfach zu Sprachverboten, die sich auch in mehr oder weniger systematischer Vernichtung des Schrifttums niederschlugen. Vgl. dazu SKL 2014, v. a. die Artikel „Bibliografie“, „Dezemberreskript“, „Sprachverbote“.

[27] Bei Vorliegen von Editionen wurden für die älteste Phase auch ausgewählte Handschriften berücksichtigt. Eine vollständige Übersicht über einzelne niedersorbische Texte in überwiegend obersorbischen Zeitschriften wurde bisher noch nicht erstellt. Bei Werken, die in mehreren Auflagen erschienen (z. B. Gesangbücher), wurden nur ausgewählte, stark veränderte Neuauflagen zusätzlich aufgenommen.

[28] Für diese Aufgabe ist Fabian Kaulfürst zuständig, ebenso wie seit einigen Jahren für die Organisation der Primärdigitalisierung. Die Informationen zur Schrifttums-Inventarisierung im folgenden Abschnitt über die Primärdigitalisierung stützen sich im Wesentlichen auf seine Angaben.

weitgehend[29] abgeschlossen, sodass mittlerweile die – statistisch gesprochen – Grundgesamtheit dessen, was durch das niedersorbische Textkorpus repräsentiert wird, für diese Zeit gut bekannt ist.

4.2 Primärdigitalisierung: Vom Druckwerk zum Volltext

Mit Primärdigitalisierung ist der Prozess der Überführung von „Text" aus analoger (z. B. eine gedruckte Zeitungsseite) in digitale Form gemeint (d. h. in elektronischen Text, auch *Volltext* genannt). Dies geschieht in der Regel entweder über (mehr oder weniger professionelles) Abschreiben mittels Computer, z. B. im Double-Keying-Verfahren, oder mittels automatischer Texterkennung (Optical Character Recognition – OCR). Ein Volltext als Ergebnis der Primärdigitalisierung gilt mit Blick auf das neue Textkorpus zunächst als Rohtext.

Beiden Verfahren ist in der Regel eine Bild-Digitalisierung vorgelagert, um den Arbeitsprozess zu erleichtern bzw. zu ermöglichen. Hochwertige Bilddigitalisate dienen außerdem u. a. der Bestandsicherung oder besseren Verfügbarkeit. Die Primärdigitalisierung bereitet beim sorbischen Schrifttum immer noch Probleme, da die Volltextqualität beim OCR älterer sorbischer Frakturtexte[30] nach wie vor unzureichend ist – ein Großteil des sorbischen Schrifttums erschien bis in die 1930er-Jahre in Frakturschrift. Nur für das Niedersorbische konnte dieses Problem kompensiert werden, indem mit Hilfe von Drittmitteln über etwa 15 Jahre eine umfängliche Volltextgenerierung per Double-Keying erfolgte. Beim deutlich umfangreicheren[31] obersorbischen Schrifttum ist diese Aufgabe nach wie vor ungelöst, und unsere Hoffnung liegt dort auf einer sich verbessernden OCR-Technologie.[32]

Im Niedersorbischen konnte das als korpusrelevant eingestufte und verfügbare Schrifttum bis 1848 bereits weitgehend und in hoher Qualität primärdigitalisiert werden.

---

[29] Dieser Vorbehalt betrifft vor allem mögliche Lücken in den Bibliografien. Es ist damit zu rechnen, dass hin und wieder gänzlich unbekannte Schriften auftauchen oder als verschollen geglaubte Drucke ermittelt werden – siehe z. B. Kaulfürst 2010. Daher ist zu erwarten, dass sich das Korpus weiterentwickeln wird. Es ist jedoch nicht mit umfänglichen Veränderungen der Zusammensetzung zu rechnen.

[30] „Die Ergebnisse der Pilotprojekte zeigen, dass bei Frakturschrift in Zeitungen, die zu den komplexesten und damit schwierigsten OCR-Vorlagen zählen, über 95 % Zeichengenauigkeit erreichbar sind, die einer anzustrebenden Wortgenauigkeit von ca. 80 % entsprechen." (Masterplan Zeitungsdigitalisierung 2017: 34). Eine Wortgenauigkeitsquote dieser Höhe bedeutet freilich, dass mindestens etwa jedes fünfte Textwort fehlerhaft digitalisiert wurde, was mit Blick auf die im Folgenden beschriebenen Aufbereitungsschritte für Volltexte einen nicht zu bewältigenden Mehraufwand verursachen würde.

[31] Für das historische Schrifttum liegt noch keine hinreichende Datenbasis vor, um zuverlässig sagen zu können, um welchen Faktor das gesamte obersorbische Schrifttum das niedersorbische übersteigt. Für das Jahr 2019, so lässt sich auf Grundlage des gerade begonnenen Schrifttums-Monitorings (vgl. Kap. 4.6) feststellen, beträgt das Verhältnis ziemlich genau 4:1. (Die zugrundeliegenden Token-Zahlen finden sich in Fußnote 75.)

[32] Hier sind vor allem die Entwicklungen im durch die Deutsche Forschungsgemeinschaft geförderten OCR-D-Projekt vielversprechend: https://ocr-d.de/de/. Dort heißt es: „Die Verfügbarmachung von Volltexten zum Zweck der Volltextsuche und Weiterbearbeitung, bspw. mit Werkzeugen der Digital Humanities, ist ein großes Desiderat der Forschung, das durch eine koordinierte Förderinitiative zu bearbeiten ist." Das Sorbische Institut bringt sich hier über Partner mit Blick auf sorbische Spezifika ein. Vgl. auch Fußnote 46.

Ein Lückenschluss erfolgt zurzeit zunächst für die Zeit bis 1945. Das folgende Diagramm stellt den derzeitigen Umfang (Stand Ende 2019) sowie die Zusammensetzung des niedersorbischen Textkorpus dar. Es umfasst gut 43 Millionen Tokens[33] in mehr als 7200 Einzeltexten. Dabei wurden KORPUSTEXTE sowie abschließend bearbeitete ROHTEXTE einbezogen (Stufe 1 lt. Abb. 3). Es handelt sich somit in erster Linie um eine Information zum Stand der Primärdigitalisierung, die von der Differenzierung zwischen „altem" und „neuem" Textkorpus abstrahiert.

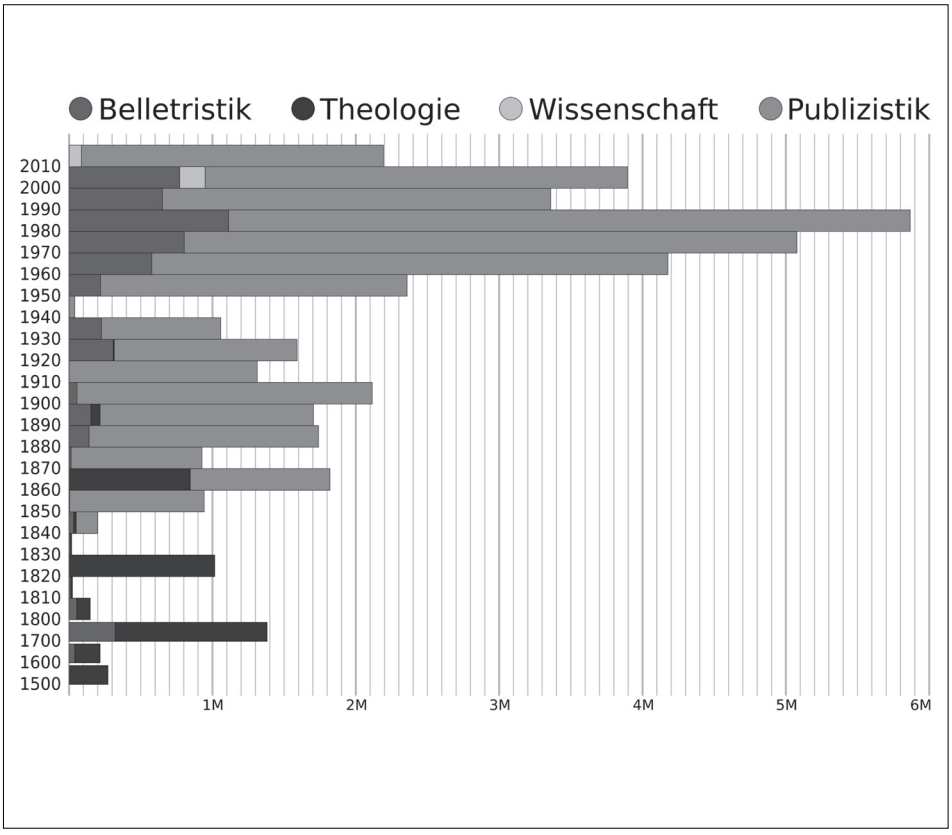


Abb. 2: Zusammensetzung des niedersorbischen Textkorpus als Abbild der Primärdigitalisierung

Grob gesprochen wird das niedersorbische Schrifttum bis Mitte des 19. Jahrhunderts von religiösen Einzeltexten (z. B. Ausgaben des Neuen und/oder Alten Testaments) dominiert. Wie in der Grafik erkennbar, setzt erst danach, vor allem mit dem Beginn publizistischen

[33] Die Zählung ist hier vorläufig, da sie der in Kap. 4.4.2 dargestellten eigentlichen Tokenisierung vorgelagert ist. So werden etwa „Tokens" mitgezählt, die später aus der weiteren Bearbeitung ausgeschlossen werden, z. B. weil es sich um Wörter deutscher Textpassagen in einem ansonsten niedersorbischen Text handelt (vgl. Kap. 4.4.3 und 4.4.6).

Schreibens seit der ersten Ausgabe der Wochenzeitung „Bramborski Serbski Casnik“[34] am 5. Juli 1848, ein zunehmend auch weltliches Schrifttum ein.

Da die Primärdigitalisierung der niedersorbischen Publizistik bis 1989 vollständig abgeschlossen ist und die Digitalisate auch bereits hinreichend bearbeitet wurden, um in die Grafik eingehen zu können (s. o.), spiegelt der Umfang der Texte im Korpus diesen Teil des Schrifttums für die Zeit zwischen 1848 und 1989 exakt wider.[35] Während sich der Umfang publizistischer Texte demnach seit den 1850er-Jahren relativ konstant zwischen knapp einer und etwa zwei Millionen Tokens pro Jahrzehnt bewegt (mit einem Höhepunkt im ersten Jahrzehnt des 20. Jahrhunderts), bricht diese Textproduktion – wie in der Grafik deutlich zu erkennen – mit dem erzwungenen Ende des BC als nach wie vor einziger Zeitung[36] – die letzte Ausgabe erschien am 29. Juli 1933 – nahezu vollständig[37] zusammen, und zwar noch vor dem umfassenden nationalsozialistischen Sprachverbot ab 1937.[38] Nach dem Zweiten Weltkrieg gelang es nicht sofort, erneut eine niedersorbische Wochenzeitung auf die Beine zu stellen. 1947/48 erschienen insgesamt nur drei Ausgaben der neuen Zeitung unter dem Titel „Dolnoserbski Casnik“ (dt. „Niedersorbische Zeitung“), und zwar als Beilage der obersorbischen Tageszeitung „Nowa doba“ (dt. „Neue Zeit“). Von 1949 bis 1954 erschien diese Beilage dann monatlich, jetzt unter dem bis heute gültigen Namen „Nowy Casnik“ (Abk. NC; dt. „Neue Zeitung“). Erst 1955 gelang es, den NC erneut als eine regelmäßig erscheinende Wochenzeitung zu etablieren.[39] Die starke quantitative Dominanz der Publizistik auch nach der Mitte des 20. Jahrhunderts dürfte sich allerdings teilweise noch etwas relativieren, da die bisherige Primärdigitalisierung für diese Zeit stark auf dieses Genre fokussiert war.[40]

---

[34] Die Zeitung erschien in den Folgejahren bis 1933 unter wechselnden Namen, wird aber in diesem Artikel davon absehend als „Bramborski Casnik“, kurz BC, bezeichnet. Der Nachfolger nach dem Zweiten Weltkrieg heißt „Nowy Casnik“ (NC).

[35] Eine zügige hochwertige Volltext-Digitalisierung auch obersorbischer Zeitungen wäre auch deshalb von besonderer Wichtigkeit für die Forschung, weil zahlreiche, vor allem historisch-kulturwissenschaftliche Fragestellungen, zu deren Behandlung die Auswertung von Textkorpora herangezogen werden kann, nur bei Berücksichtigung beider sorbischer Schriftsprachen beantwortet werden können.

[36] Daneben existierten sporadisch niedersorbische religiöse Zeitschriften, vor allem der zwischen 1904 und 1913 gedruckte „Wosadnik“ (dt. „Gemeindebote“). Dieser ist mit einem Umfang von etwa 335 000 Tokens ebenfalls Bestandteil des Textkorpus.

[37] Bis 1937 erschienen nur noch wenige kleinere Publikationen (von Bogumił Šwjela, Mina Witkojc und Marjana Domaškojc) sowie einige kurze Einzeltexte, insgesamt weniger als 250 Druckseiten.

[38] Schon von April 1918 bis Ende 1920 hatte es bedingt durch den Ersten Weltkrieg und seine Folgen eine Unterbrechung gegeben, die in der Grafik wegen der jahrzehnteweisen Darstellung nicht erkennbar ist. 1921 erschien die Zeitung erneut, jedoch nicht sofort wöchentlich.

[39] Eine kurze Geschichte der Zeitung (auf Niedersorbisch) von Gregor Wieczorek, Chefredakteur bis April 2020, findet sich auf der Internetseite des NC: https://www.nowycasnik.de/index.php/dsb/historija-nowego-casnika [06.03.2020].

[40] Dem zuvor Gesagten entsprechend spiegelt die in Abb. 2 erkennbare Abnahme des Umfangs publizistischer Texte ab 1990 keine Veränderung im Schrifttum wider, sondern lediglich den noch nicht abgeschlossenen Ausbau des Textkorpus.

4.3 Textaufbereitung I: Vom ROHTEXT zum KORPUSTEXT

Die aus der Primärdigitalisierung hervorgehenden ROHTEXTE werden durch ein mehrstufiges Verfahren zur Textaufbereitung und Qualitätssicherung weiter verarbeitet zu KORPUSTEXTEN:

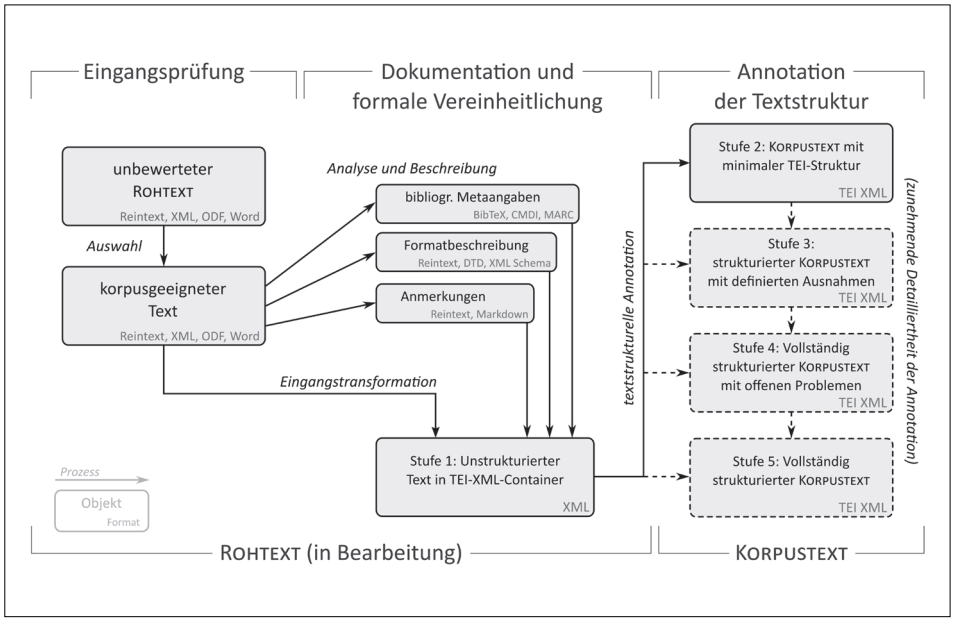


Abb. 3: Arbeitsschritte auf dem Weg vom ROHTEXT zum KORPUSTEXT

4.3.1 Prüfung der Korpuseignung

Wahrscheinlich wird man in vielen Vorhaben zum Aufbau von Textkorpora die Erfahrung gemacht haben, dass (oft lang ersehnte) von Einzelpersonen nach mühsamer Abschrift gelieferte, einstmals per OCR digitalisierte oder sogar von Verlagen zur Verfügung gestellte Volltexte (aus der Korpusperspektive) von so zweifelhafter Qualität sind, dass sie nicht ins Textkorpus aufgenommen werden können. Gründe dafür sind vielfältig: Der Text wurde nicht normgerecht und/oder inkonsistent normalisiert, die Fehlerrate beim OCR war unangemessen hoch, der Text entspricht einer Fassung vor der Endredaktion und daher nicht dem Drucktext. Oftmals ist eine nachträgliche Korrektur unmöglich oder unangemessen aufwendig. Daher ist die erste Hürde vom ROHTEXT zum KORPUSTEXT eine Prüfung seiner generellen E i g n u n g als Bestandteil des anvisierten Textkorpus, wie der erste Bearbeitungsschritt in Abb. 3 darstellt. Dabei können neben der formalen Qualität[41] eines Volltext-Digitalisats auch inhaltliche oder „normative“ Kriterien eine

[41] Die Qualitätskriterien für KORPUSTEXTE im Sinne des neuen Korpus wurden deutlich verschärft, sodass faktisch nur noch durch das professionelle Double-Keying-Verfahren digitalisierte Texte mit einer Buchstabengenauigkeit von mindestens 99,95 % – so die von uns aufgrund regelmäßiger Stichproben als zutreffend eingestuften Leistungszusagen des

Rolle spielen (schriftsprachlicher oder dialektaler Text, Grad der Bearbeitung bei Neuauflagen, Zuverlässigkeit der Quelle bei nicht durch anerkannte Instanzen redigierten Texten, Vollständigkeit usw.).

4.3.2 Metadaten und formale Vereinheitlichung

Nach dieser ersten Hürde, dargestellt in der Grafik-Spalte „Eingangsprüfung", wird ein als korpusgeeignet bewerteter ROHTEXT durch die Erfassung seiner Eigenschaften sowie eine formelle Vereinheitlichung weiterverarbeitet:

1. Beschreibung des Textes hinsichtlich (a) der notwendigen bibliografischen Angaben, (b) der vorliegenden formalen Eigenschaften (z. B. Zeichenkodierung) und (c) weiterer relevanter Angaben (z. B. Herkunft aus einfacher manueller Abschrift, aus professionellem Double-Keying oder einem OCR-Verfahren). Diese Angaben dienen zugleich der
2. Unifizierung der Zeichenkodierung. Außerdem wird
3. der potenzielle KORPUSTEXT mittels sehr basaler struktureller Annotation in einen TEI-XML-Container überführt, der dem TEI-Schema[42] zunächst nur formell genügt.

4.3.3 Textstrukturelle Annotation

Neben der generellen Korpuseignung und der soeben beschriebenen grundlegenden formalen Vereinheitlichung wird für die sorbischen Textkorpora als drittes Kriterium der KORPUSREIFE eine minimale textstrukturelle Annotation im TEI-Format gefordert (Stufe 2 lt. Abb. 3). Zu diesem Zweck wurde ein TEI-konformes Annotationsformat für KORPUSTEXTE definiert, das zunächst einer schwerpunktmäßig inhaltlich-textsemantisch angemessenen strukturellen Modellierung dient, aber darüber hinaus auch die Informationen weitergehender Annotationsschritte (vgl. Kap. 4.4) aufnimmt bzw. aufnehmen wird.[43]

Dienstleisters – aufgenommen werden können. Teilweise mussten daher nach der Eingangsprüfung alte Korpustexte erneut digitalisiert werden. Im Zuge dieser Neubewertung als nicht korpusgeeignet ausgesonderte ROHTEXTE stammten aus zwei Quellen: Zum einen aus einer sehr frühen Phase manueller Volltextdigitalisierung. In diesen Abschriften durch Hilfskräfte vom Ende der 1990er-Jahre finden sich nicht nur zahlreiche Abschreibfehler, sondern es wurde zudem häufig „intuitiv", unsystematisch und undokumentiert normalisiert (orthografisch, teilweise aber auch morphologisch), sodass diese Texte kein getreues Bild des Originaltextes bieten und die fehlerhaften Normalisierungen auch nicht rückgängig gemacht werden können. Zum anderen, wegen des Fehlens von geeigneter Technik wie auch von Finanzmitteln für Double-Keying-Dienstleistungen, aus einer versuchsweise in Kooperation mit der Diakonie Niederlausitz durchgeführten AB-Maßnahme (2004–2008) zum quasi unprofessionellen „Single-Keying".

[42] Die Wahl des TEI-P5-Formats entspricht den DFG-Praxisregeln „Digitalisierung" nach DFG-Vordruck 12.151 vom Dezember 2016, Kap. 3.3.2 (Strukturelle Metadaten für digitale Faksimiles) sowie den Empfehlungen von CLARIN-D (CLARIN-D User Guide, Version 1.0.1, 2012-12-1). Die TEI-P5-Guidelines finden sich unter http://www.tei-c.org/guidelines/p5/.

[43] Die zu diesem Zweck an der Cottbuser Zweigstelle des SI ab 2016 definierte Teilmenge des TEI-P5-Formats unter der Bezeichnung „Cottbus Corpus Format (CCF)" wurde von Marcin

Die im Rahmen dieses Gesamtmodells als „Stufe 2“ (von insgesamt fünf Stufen bis zur vollständigen textstrukturellen Ausmodellierung, z. B. einschließlich komplexer Tabellen) gekennzeichnete textstrukturelle Annotation zeichnet den Text soweit aus, dass er korpusanalytisch effektiv nutzbar ist und Text-/Fundstellen sich präzise identifizieren und indizieren lassen. Außerdem ist eine basale Textpräsentation möglich.[44] Diese Stufe der textstrukturellen Annotation ermöglicht somit ein vertretbares Verhältnis von Aufwand und Nutzen.[45] Ausgezeichnet wird auf dieser Stufe die Grobstruktur des Textes (`text/front, text/body, text/back`), einzelne Texte mit Titel/Autor (z. B. Artikel in Zeitungen – `div[@type='article']` –, Einzeltexte eines Sammelbandes u. ä.) werden markiert und ggf. vorhandene Titel (`head`) bzw. Autoren annotiert (`byline`), Bilder und Bildunterschriften werden gekennzeichnet (`figure`) sowie Angaben zur Seitennummerierung und zum Textfluss ergänzt. Diese textstrukturelle Standard-Annotation für KORPUSTEXTE wird fakultativ erweitert: So werden beispielsweise fast alle belletristischen Texte sowie die meisten Ausgaben des niedersorbischen Jahreskalenders „Pratyja“ bis Stufe 4 oder 5 annotiert.

Das oben erwähnte Ziel einer Qualitätssicherung mit Blick auf alle KORPUSTEXTE bleibt über alle Bearbeitungsstufen hinweg relevant, besonders auch noch im Zuge der textstrukturellen Annotation und hier insbesondere zwischen Stufe 1 und 2. Denn bei der XML-TEI-Modellierung zur Stufe 2 (oder höher), d. h. wenn auch Überschriften, Absätze usw. ausgezeichnet werden, wird der Text zwangsläufig genauer geprüft und ggf. auch noch einmal mit dem analogen Original verglichen. Spätestens hier fallen diverse Typen von Fehlern auf, die bis zu dieser Bearbeitungsstufe nicht bemerkt wurden. Im Verlauf der Bearbeitung wird quasi in den Text gezoomt, von der Gesamtschau bis auf die Seiten/Absatz-Ebene bei der textstrukturellen Annotation und bis auf die Wortebene in einer noch folgenden Bearbeitungsstufe.[46]

Szczepański zusammengestellt und dokumentiert und gemeinsam mit Joanna Szczepańska an sorbischen KORPUSTEXTEN getestet und optimiert. Sie wird stetig weiterentwickelt und z. B. um Elemente der linguistischen Analyse ergänzt (vgl. Kap. 4.4). Der derzeitige Stand ist dokumentiert auf https://niedersorbisch.de/korpus/format/. Im Unterschied zum DTA-Basisformat (s. http://www.deutschestextarchiv.de/doku/basisformat/) fordert das CCF keine so starke formale Originaltreue, die für unsere Zwecke verzichtbar war. Das CCF zielt prioritär auf eine semantisch-funktionale Modellierung und vermeidet diverse obligatorische Format-Auszeichnungen, deren Beachtung die uns verfügbaren Kapazitäten stark strapaziert hätte (vgl. auch Kap. 5.2.3). Außerdem nutzt das CCF zusätzlich TEI-Elemente für die linguistische Analyse mit gewissen eigenen Erweiterungen (vgl. TEI unter https://www.tei-c.org/release/doc/tei-p5-doc/en/html/AI.html#AILALW). So dient es auch als Grundlage für eine linguistische Annotation der KORPUSTEXTE (vgl. Kap. 4.4).

44 So sind die Annotationen dieser Stufe weitgehend hinreichend für eine Präsentation der relativ einfach strukturierten Ausgaben des „Bramborski Casnik“ (s. Kap. 4.2), nicht aber von komplexer strukturierten Texten. Dort erfordert eine annähernd vollständige und relativ originalgetreue Textpräsentation eine höhere Annotationsstufe. Dies betrifft etwa die Ausgaben der „Pratyja“, wie oben im Folgetext erwähnt.

45 Auch die Herstellung dieser textstrukturellen Annotationsstufe ist bereits mit einigem Aufwand verbunden, sodass in Zukunft und bei größeren Textmengen (z. B. obersorbischer Zeitungstexte; siehe aber auch die folgende Fußnote) für Teile des GESAMTKORPUS auch ein vorläufiger Verbleib in Stufe 1, d. h. in einem einfachen TEI-Container erwogen werden muss, was für bestimmte Nutzungen ausreichend ist. Insofern stellt die Stufe-1-Bearbeitung eine Art „Wartezone“ vor der KORPUSREIFE dar. Ausschlaggebend sind hier die für diesen Arbeitsschritt verfügbaren Ressourcen.

46 Es ist möglich, dass die in der dritten Spalte von Abb. 3 dargestellte textstrukturelle Annota-

4.4 Textaufbereitung II: Analyse und Annotationen für ein Globalkorpus

Angenommen, es lägen vom gesamten publizistischen Schrifttum des Niedersorbischen hochwertige Korpustexte vor – ein Zustand, der erfreulicherweise in den nächsten Jahren erreicht werden dürfte. Damit stünde für diesen Teil des Schrifttums ein historisches Vollkorpus für verschiedene Arten der Auswertung zur Verfügung. Sein großer Vorteil wäre die textuelle Zuverlässigkeit, sein größter Nachteil das Fehlen von über den reinen Text und dessen textuelle Struktur (Seiten, Überschriften, Absätze usw.) hinausgehenden Annotationen. Alle Suchanfragen müssten sich bei solch einem Korpus ausschließlich an konkreten Textwörtern (Tokens) orientieren. Was das bedeutet, soll im Folgenden verdeutlicht werden.

Zum einen sind Textwörter in einer flektierenden Sprache wie dem Niedersorbischen in aller Regel Flexionsformen (z. B. *bokoju* als Dativ Singular von ns. *bok* ‚Seite'). Manchmal werden auch genau diese gesucht, häufiger jedoch, vor allem in nicht-linguistischen Abfragen wegen des stärker inhaltlichen Interesses, richtet sich die Suche auf ein „Wort" im Sinne eines Lexems oder „Begriffs" – u. a. im Kontext der Korpusaufbereitung spricht man auch von Lemmata – als Oberbegriff für alle[47] zugehörigen Flexionsformen. Eine solche Suche muss daher anstelle eines einfachen Suchbegriffs (z. B. *bokoju*) eine komplette Liste sämtlicher einem gesuchten Lexem (oder noch umfangreicher: einer Lexem-Kombination) zugeordneter Wortformen umfassen. Dies verlangt jedoch neben Sprachkompetenz im Sinne von Textverstehen bereits aktive Kenntnisse des grammatischen Systems und verursacht außerdem einen erheblichen Mehraufwand.[48] Während man von Linguisten erwarten kann, dass alle oder zumindest die meisten Flexionsformen ermittelt und damit bei einer Suche erfasst werden (auch mit technischen Hilfsmitteln), dürfte dies für Nicht-Linguisten eine nicht so einfach überwindbare Hürde darstellen.

Zum anderen handelt es sich beim niedersorbischen (publizistischen) Korpus um ein historisches. Wir haben es daher in den Texten aus verschiedenen Entstehungsperioden – bei einer wie dem Niedersorbischen relativ schwach normierten Sprache sogar auf synchroner Ebene – mit einer erheblichen Variation der Schreibung „derselben" Wortformen zu tun. Das tatsächliche Maß der Schreibvariation, das für eine zuverlässige Suche und Datenauswertung berücksichtigt werden müsste, kann sehr unterschiedlich ausfallen.[49] Diese Beurteilung erfor-

---

tion und die damit verbundenen Maßnahmen zur Qualitätskontrolle bei Teilen des obersorbischen Textkorpus modifiziert oder gänzlich anders erfolgen werden. Dies wäre damit zu begründen, dass dort – wie in Kap. 4.2 dargestellt – voraussichtlich ein Großteil der Rohtexte über eine stetig verbesserte OCR erzeugt werden wird. Da mittlerweile prinzipiell auch das „Layout" automatisch erkannt und markiert werden kann (Optical Layout Recognition, OLR), könnte eine manuelle textstrukturelle Annotation (zumindest zum Teil) hinfällig sein. Vgl. auch Fußnote 32.

47 Nur einen Teil zu erfassen, würde – da dann nicht alles Relevante berücksichtigt (und damit gezählt; vgl. Kap. 3) werden kann – die Abfrageergebnisse verfälschen und damit ihre Validität infrage stellen.

48 Bei einem niedersorbischen Substantiv (z. B. *bok* ‚Seite') müssen bei sechs Kasus und drei Numeri jeweils maximal 18 Flexionsformen berücksichtigt werden (im konkreten Fall wegen homonymer Formen 13: *boce, bok, boka, bokach, bokam, bokami, boki, bokoju, bokom, bokoma, bokow, bokowu, boku*); bei anderen Wortarten kann die Anzahl noch höher sein.

49 Einen extremen, aber keineswegs unrealistischen Fall stellen mögliche Schreibvarianten der niedersorbischen Äquivalente für dt. *alle* dar: allein die Pluralform *wšykne* zeigt 48 mögliche Schreibungen: *schikne, šykne, fchickne, fchikne, fchyckne, fchykne, fchykne, wšikne,*

dert aber bereits gründliche sprachhistorische Kenntnisse. Eine diesbezüglich unreflektierte Suche bedeutet, den Zufall „ins Spiel“ zu bringen, was jedenfalls kein wissenschaftliches Vorgehen wäre. Von kontrollierter Zuverlässigkeit könnte keineswegs mehr die Rede sein.

Diese beiden Probleme beim Zugang zu KORPUSTEXTEN, die im Sinne der am Anfang von Kapitel 3 genannten Handlungsmaximen als Nutzungsbarriere wirken, werden durch zwei Bearbeitungsschritte behoben, die als Normalisierung und Lemmatisierung bezeichnet werden. Diese werden im Folgenden, in Zusammenhang mit weiteren notwendigen und in der Korpuslinguistik gängigen Aufbereitungsschritten, der Reihe nach dargestellt. Einen Überblick gibt zunächst die folgende Abbildung:

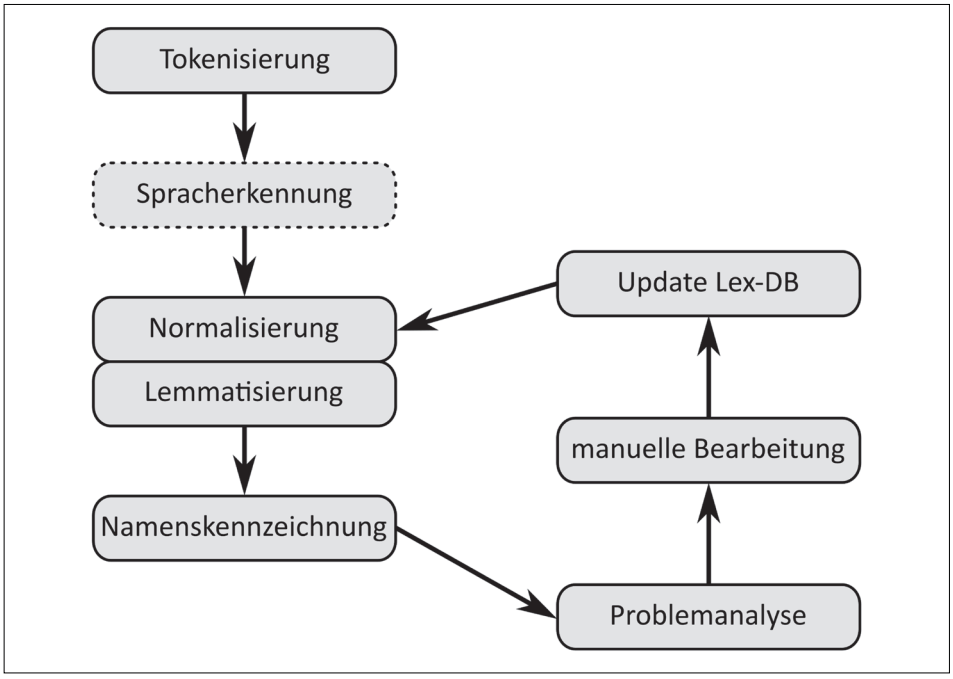


Abb. 4: Übersicht der Arbeitsschritte zur Aufbereitung von KORPUSTEXTEN

Das darzustellende Verfahren zur Aufbereitung von KORPUSTEXTEN ist kontrolliert, kollaborativ organisiert und bezieht teilautomatisierte Prozesse mit ein. Als organisatorische Klammer dient eine 2016–2019 in einem Drittmittelprojekt konzipierte und implementierte Software namens Corproc (s. Kap. 4.4.1). Eine entscheidende Rolle im Gesamtkon-

wfchykne, wfchykne, usw. usf. Der diese Formen erfassende Reguläre Ausdruck wäre /w?(š|(s|ſ|ʃ)ch)(i|y)(c|k)?kne/. Denselben Wortbildungsstamm zeigen aber auch die fünf niedersorbischen Äquivalente der deutschen Wörter *jeder/jede/jedes* (*wšyken/wšykna/wšykno/wšyknej/wšykne*) zzgl. der Flexionsformen. Die Einbeziehung der dort möglichen Schreibvarianten, repräsentiert durch den Regulären Ausdruck /w?(š|(s|ſ|ʃ)ch)(i|y)(c|k)?k(en|n(a|e(n|(j|y|i)(e|u)?)?|o(go?|mu?)?|u|(y|i)(ch|m(i|y)?)))/, ergibt 1152 mögliche Schreibungen, und zwar noch ohne Anspruch auf Vollständigkeit. (Ich danke Marek Slodička für dieses schlagende Beispiel, das bereits Gegenstand des in Fußnote 3 erwähnten Vortrags war.)

zept spielt außerdem die bereits in Kap. 1 erwähnte lexikalische Datenbank (Lex-DB)[50], die – bestehend aus verschiedenen Modulen – vor allem folgende Informationen enthält:

- Eine Liste von – grob gesprochen – „Lemmata" mit sämtlichen zuordenbaren (gebuchten und/oder generierten) Flexionsformen; für einen Teilbestand zusätzlich die Kennzeichnung, dass es sich um einen Eigennamen handelt (vgl. Kap. 4.4.4; in den Folgekapiteln wird diese Liste auch als „Vollformenlexikon" bezeichnet);
- eine Liste registrierter Abkürzungen sowie
- eine Liste zusätzlicher (überwiegend nicht-sorbischer, aber auch aus anderen Gründen bisher im Vollformenlexikon nicht erfasster) Eigennamen.

Wie in Abb. 4 deutlich wird, ist die Lex-DB einerseits eine wichtige Grundlage für die automatisierte Normalisierung und Lemmatisierung, wird aber gleichzeitig im Gesamtprozess ständig weiter angereichert. Das langfristige Ziel dieses iterativen „Abgleichs" der Lex-DB mit neuen Korpustexten ist eine nahezu vollständige maschinenlesbare Erfassung des im niedersorbischen Schrifttum belegten Wortschatzes.[51]

4.4.1 Corproc[52]

Corproc steht für *corpus processing* und ist der Name einer webbasierten Arbeitsumgebung zur Verarbeitung und Anreicherung von TEI-konformen Korpustexten. Sie wurde mit gängigen und nachhaltigen Webtechnologien umgesetzt (HTML, CSS, JavaScript, PHP). Die Arbeitsumgebung zeichnet sich durch folgende Merkmale aus:

- einfache Bedienung auch ohne informatische oder computer- und korpuslinguistische Kenntnisse
- keine (über einen Webbrowser hinausgehende) Installation zusätzlicher Software
- geringe Anfälligkeit für menschliche Fehler

---

[50] Diese Datenbank geht zurück auf eine erste Fassung, die seit 2014 unter Federführung von Marcin Szczepański in Zusammenhang mit der Entwicklung einer automatischen Rechtschreibprüfung für Niedersorbisch und eines dieser zugrunde liegenden Wortformen-Generators entstand – damals (2014/15) gefördert von der Vattenfall Europe Mining AG (im Rahmen einer Vereinbarung mit der Domowina, Regionalverband Niederlausitz), und kofinanziert durch das Sorbische Institut und das WITAJ-Sprachzentrum (RCW). Diese Datenbasis wurde dann in den Folgejahren stetig erweitert, u. a. ab 2017 in Drittmittelprojekten zur Weiterentwicklung der automatischen Rechtschreibkontrolle, die von der Stiftung für das sorbische Volk im Rahmen der Förderinitiative „Sorbisch in den neuen elektronischen Medien" gefördert wurden, gespeist aus Sondermitteln des Bundes und der Länder Sachsen und Brandenburg.

[51] Diese könnte dann u. a. als Grundlage für eine umfassende lexikografische Beschreibung des historisch belegten niedersorbischen Wortschatzes dienen (vgl. Bartels 2013).

[52] Das Programm wurde entwickelt im Rahmen des Drittmittelprojekts „Sorbenwissen", das von 2016–2019 gemeinsam vom Sorbischen Institut und der Technischen Universität Dresden durchgeführt und über den Freistaat Sachsen durch den Europäischen Sozialfonds gefördert wurde. Corproc wurde – in enger Abstimmung mit den in der Cottbuser Zweigstelle am Textkorpus Arbeitenden – von Marek Slodička konzipiert und implementiert. In die Textabschnitte 4.4.1 bis 4.5 ist eine textuelle Zuarbeit von M. Slodička eingeflossen.

- Unterstützung verschiedener Objektsprachen (derzeit: Niedersorbisch und Obersorbisch)
- Modularität, Fakultativität und Iterierbarkeit einzelner Verarbeitungsschritte
- simultane Verarbeitung verschiedener KORPUSTEXTE
- sukzessiv-kollaborative Verarbeitung einzelner KORPUSTEXTE durch verschiedene Personen mithilfe einer einfachen Benutzer*innen- und Rollenverwaltung

Corproc vereint unterschiedliche Schritte der Verarbeitung von KORPUSTEXTEN zu einem umfassenden und standardisierten Workflow mit einer auf zehn Rollen[53] genau differenzierten und scharf abgegrenzten Zuständigkeitsdistribution und erleichtert die Annotation von KORPUSTEXTEN und deren Qualitätsmanagement. Die durch Corproc gesteuerten Verarbeitungsschritte werden in den folgenden Abschnitten beschrieben.

4.4.2 Tokenisierung

Unter Tokenisierung wird die Segmentierung eines Fließtextes in einzelne maschinenlesbare Einheiten (Tokens) verstanden (SCHMIDT 2008, HAGENBRUCH 2010; vgl. auch Fußnote 9). Die Tokenisierung bildet die Grundlage aller nachfolgenden Verarbeitungsschritte, da diese nicht mit dem jeweiligen Text als Ganzes, sondern mit der Gesamtheit der einzelnen Tokens arbeiten. Bei der Tokenisierung mit Corproc können neben Satz- und Leerzeichen auch Abkürzungen berücksichtigt werden sowie sprach- und epochentypische Abweichungen in der Schreibung grafischer Wörter von der synchronen usuellen Tokenstruktur. Worttokens werden TEI-konform mit `<w>`-Elementen, Interpunktionstokens mit `<pc>`-Elementen umgeben, wodurch diese Gliederung sowie die einzelnen Tokens für einen kontrollierten Zugriff durch den Computer vorbereitet werden. Ein einfaches Beispiel zur Illustration:

```
<w>Tokenizacija</w> <w>jo</w> <w>ważny</w> <w>kšac</w>
<w>pśigótowanja</w> <w>korpusowych</w>  <w>tekstow</w><pc>.</pc>
```[54]

Bei der Tokenisierung tritt bekanntermaßen eine ganze Reihe (mehr oder weniger) sprachspezifischer Probleme auf, so u. a. die Mehrfunktionalität von Interpunktionszeichen (z. B. der Punkt am Ende eines Satzes oder einer Abkürzung, die aber selbst auch am Ende eines Satzes stehen kann), die Wortverbindungen mit Bindestrich (*śěg Barliń-Warszawa, Schuster-Šewc, buźomy-li*) oder Schrägstrich (*Mucke/Muka, wón/a, Cottbus/Chóśebuz*) – handelt es sich jeweils um ein Token oder um mehrere? Zusätzliche Probleme entstehen häufig im Zuge der Digitalisierung, z. B. dadurch, dass in der Druckfassung enthaltene Worttrennungen (*tek-stow*) nicht entfernt wurden oder dass Wörter fälschlich zusammengerückt werden oder auch im Originaltext gelegentlich zusammengeschrieben werden (*ktomu*).

Für den in diesem Artikel beschriebenen Aufbereitungsprozess wurde ein eigener Tokenisierer programmiert[55], der im Verfahrensverlauf weiterentwickelt wird. Hierfür war u. a. die Notwendigkeit entscheidend, die Tokenisierung auf Basis von TEI-XML-Do-

---

[53] Zu den definierten Rollen und zum entsprechenden Workflow s. Kap. 4.5.
[54] Deutsch: Tokenisierung ist ein wichtiger Schritt der Vorbereitung von KORPUSTEXTEN.
[55] Dies wurde und wird von Marcin Szczepański verantwortet.

kumenten vornehmen zu können. Wichtig für die Weiterentwicklung sind die bei der Textanalyse auftretenden Probleme (vgl. Kap. 4.4.5), die z. B. zur Ergänzung von Funktionen oder spezifischen Wortlisten führen. Dies betrifft etwa die automatische Trennung bestimmter Zusammenschreibungen in alten Texten (*steje* statt *s teje, ktomu* statt *k tomu*, wobei die Originalschreibung selbstverständlich erhalten bleibt), die reguläre Annotation von Schreibungen mit einzelnem Schrägstrich als getrennte Tokens (*Mucke/Muka* → `<w>Mucke</w><pc>/</pc><w>Muka</w>`) oder die automatische Abtrennung der Partikel *-li/-lic*.

4.4.3 Normalisierung und Lemmatisierung

Nach dieser Vorbereitung folgen angesichts der zu Beginn des Kapitels geschilderten Probleme mit Flexionsformen und ihren (historischen) Schreibvarianten die wichtigsten Aufbereitungsschritte der Normalisierung und Lemmatisierung.

Mit Normalisierung wird in der Korpuslinguistik generell ein Analyse- und Annotationsvorgang bezeichnet, „bei dem variierende Formen mit derselben Bedeutung vereinheitlicht werden, um sie später einheitlich finden zu können." (Hirschmann 2019: 93) Häufig, so auch in unserem Fall, erfolgt eine Vereinheitlichung zur aktuellen Standardorthografie. Dabei überschreibt die normalisierte Form nicht die im Korpustext enthaltene Originalform, sondern reichert als zusätzliche Information das Token an.

Für möglichst zuverlässige und wenig übergeneralisierende Ergebnisse ist die Normalisierung sensitiv gegenüber dem Erstellungszeitraum eines Korpustextes im Rahmen von definierten und in Corproc als Parameter einstellbaren „Orthografieperioden".[56] Die Überbrückung der Differenz zwischen heutiger Standardorthografie und dem Usus der gewählten Periode geschieht durch eine Abfolge von Ersetzungen diachroner mit synchronen Zeichenfolgen, mithilfe derer die Tokens in Reguläre Ausdrücke umgewandelt werden, die dann wiederum auf ein Vollformenlexikon angewendet werden.

Die Suchtreffer für jedes Token werden dann als normalisierte Formen in einem `cpc:norm`-Attribut in einem eigenen Corproc-Namensraum des Tokens gespeichert.[57] Ist der Korpustext vorab mit `<foreign>`-Elementen oder `xml:lang`-Attributen angereichert worden, können sprachlich von der Objektsprache abweichende Passagen (vgl. Kap. 4.4.6) optional von der Normalisierung ausgenommen werden.

Um auch hier ein Beispiel anzuführen: Die Flexionsform Lokativ Singular des Wortes/Lemmas *město* ‚Stadt, Ort' lautet in heutiger Orthografie *měśe* ‚in der Stadt'. Im „Bramborski Casnik" des Jahres 1875 findet sich stattdessen noch folgende Schreibung (im Korpustext als lateinische Transliteration der originalen Frakturschrift der Druckfassung): *mėſcże*. Die notwendige Annotation zur Normalisierung, die die Suche nach

---

[56] An der Erstellung der hierfür notwendigen Grundlagen war Fabian Kaulfürst maßgeblich beteiligt. Hierbei handelt es sich zunächst um den Versuch einer zeitlichen Gruppierung bestimmter usueller wie auffälliger orthografischer Phänomene unter Berücksichtigung von Sonderentwicklungen in einzelnen wichtigen Sprachdenkmälern.

[57] Dies stellt eine geringfügige Abweichung vom TEI-Usus dar, dessen Bestimmungen eine verschachtelte, für den konkreten Einsatz wenig praktikable und unnötig redundante Auszeichnung mit den Elementen `<choice>`, `<orig>` und `<reg>` vorsehen: `<choice><orig><w>mėſcże</w></orig><reg><w>měśće</w></reg></choice>`.

dieser Wortform (Type) offensichtlich schon ziemlich vereinfachen kann, sieht dann wie folgt aus:

```
<w cpc:norm="měśće">mėĺcże</w>
```

Im nächsten Schritt, der Lemmatisierung (FITSCHEN/GUPTA 2008), werden nun die ggf. bereits normalisierten Tokens, bei denen es sich in der Mehrzahl der Fälle – wie auch im obigen Beispiel *měśće* – um Flexionsformen handelt, über einen Abgleich mit dem Vollformenlexikon einem Lemma zugeordnet. Dadurch wird ermöglicht, dass mit der Suche nach einem Lemma alle zugeordneten Flexionsformen gefunden werden.

Das Lemma wird TEI-konform im `lemma`-Attribut des jeweiligen Tokens gespeichert. Scheitert die Lemmatisierung einzelner Tokens, so werden sie daraufhin geprüft, ob eine Abkürzung vorliegt oder das Token ausschließlich aus Ziffern oder symbolartigen Zeichen besteht. Dies wird dann in einem gesonderten `type`-Attribut des Tokens entsprechend kategorisiert. Auch hier gilt: Ist der KORPUSTEXT vorab mit `<foreign>`-Elementen oder `xml:lang`-Attributen angereichert worden, können sprachlich von der Objektsprache abweichende Passagen optional von der Lemmatisierung ausgenommen werden (vgl. Kap. 4.4.6). Und auch hier überschreibt das Lemma die Originalform nicht, sondern reichert als zusätzliche Information das Token an:

```
<w cpc:norm="měśće" lemma="město">mėĺcże</w>
```

Angesichts der dargestellten Problematik und mit Blick auf die in Kap. 3 formulierten Ziele und Handlungsmaximen lässt sich festhalten: Bei historischen Korpora wie dem niedersorbischen – und perspektivisch ebenso dem obersorbischen – sind zuverlässige Analysen ohne eine vorbereitende Lemmatisierung und Normalisierung nicht möglich. Zwar verursacht eine solche Aufbereitung von Volltexten einen erheblichen Arbeitsaufwand.[58] Jedoch werden nur auf diese Weise sorbische Texte effektiv für die Forschung und vielfältige andere Recherchen zugänglich.

4.4.4 Eigennamen

Eigennamen (Ortsnamen, Personennamen usw.) spielen sowohl bei der Korpusaufbereitung als auch beim Zugriff auf Korpusdaten (vgl. Kap. 5.2) eine besondere Rolle (CARSTENSEN et al. 2010: 596 ff.). Die bekannte Problematik einer zuverlässigen Eigennamenerkennung (Named Entity Recognition) wird im Sorbischen noch dadurch verschärft, dass für andere Sprachen entwickelte sprachtechnologische Ressourcen nur sehr bedingt übertragbar sind[59] und für das Sorbische selbst kaum Vorarbeiten vorliegen. Eine nachhaltige Erkennung und Kennzeichnung von Eigennamen in den KORPUSTEXTEN ist aber schon deshalb von großer Bedeutung, um den Arbeitsaufwand bei der Analyse[60] zu

---

[58] Vor allem im Niedersorbischen ist der Aufwand für diese Aufgabe durch eine nur schwache Normierung in der Vergangenheit und die dadurch bedingte starke Schreibvarianz sehr hoch. Hinzu kommt eine bisher nur unvollständige Beschreibung des Sprachsystems.

[59] Einzelne Versuche hierzu laufen oder sind geplant. Greifbare positive Ergebnisse liegen noch nicht vor.

[60] Der Anteil von Eigennamen an nicht automatisch erkannter Lexik ist sehr hoch. Beim 2019 begonnenen Schrifttumsmonitoring (vgl. Kap. 4.6) waren unter den 8 % zunächst nicht

reduzieren – vgl. das Folgekapitel zur „Problembehandlung“. Außerdem kann nur so ein möglichst weitgehender Zugriff auf die Texte auch über Eigennamen ermöglicht werden, was wichtig ist, weil Informationen in Texten häufig über Namen, vor allem Orts- und Personennamen, gesucht werden.

In Ermangelung klarer Alternativen wird von uns auch bei den Eigennamen ein lexikonbasierter Ansatz schrittweise umgesetzt. Das bedeutet, dass in der Lex-DB enthaltene Eigennamen als solche gekennzeichnet werden, sodass diese Information im Annotationsverfahren ergänzt werden kann. Nicht in der Lex-DB enthaltene Namen werden über eine gesonderte Liste einbezogen (vgl. oben am Anfang von Kap. 4.4). Enthalten das Vollformenlexikon bzw. ergänzende Listen Angaben zum Eigennamenstatus der Lexeme, wird dieser vorläufig im `type`-Attribut des Tokens festgehalten.

```
<w lemma="Barlinaŕ" type="name">Barlinarjom</w>
```
[61]

Die Kennzeichnung von Eigennamen-Tokens in KORPUSTEXTEN erfolgt zurzeit noch undifferenziert mit dem Wert `"name"` im Attribut `type`. In späteren Schritten wird eine genauere Kennzeichnung nach verschiedenen Eigennamentypen angestrebt. Außerdem beschränkt sich die Annotation derzeit auf einzelne Tokens, berücksichtigt also noch keine Mehrwort-Eigennamen als Ganze (*Arnošt Muka, Běła Wóda, Nowa Wjas*). Auch dies wird in Zukunft anders sein: mehrere Tokens `<w></w>` werden dann voraussichtlich vom TEI-Element `<name/>` umschlossen und so als Mehr-Token-Name gekennzeichnet und können so mit zusätzlichen Informationen (wie z. B. GND-Daten oder Geo-Referenzen) angereichert werden, welche wiederum neue Nutzungsszenarien eröffnen. Zur Erkennung von mehrteiligen Eigennamen könnten auch kontextsensitive, syntaktische Verfahren Anwendung finden. Generell befindet sich dieser Teil des Systems noch in der Entwicklung.[62]

4.4.5 Problembehandlung

Neben der technisch-organisatorischen Klammerung der bisher beschriebenen Aufbereitungsschritte in einem komplexen Gesamtworkflow (s. Kap. 4.5) besteht die zentrale Funktion von Corproc darin, eine systematische und kontrollierte Bearbeitung der Fälle zu ermöglichen, die nicht – auf Basis der erarbeiteten lexikalischen Ressourcen – „automatisch“ behandelt werden konnten. In der beschriebenen lexikonbasierten Annotation von Tokens in bisher nicht analysierten KORPUSTEXTEN werden mittlerweile je nach

---

erkannten Tokens fast 15 000 Formen von Eigennamen. Hier kann durch einen entsprechenden schrittweisen Ausbau der Lex-DB in Zukunft der Aufwand für Nachbearbeitungen erheblich reduziert werden.

[61] Das Beispiel zeigt zugleich den das historische Schrifttum dokumentierenden Charakter des Auswertungsverfahrens, da hier eine in den älteren KORPUSTEXTEN frequente morphologische Nebenform zum heutigen schriftsprachlichen Lemma *Barlinjaŕ* (mit palatalem *n*) als Lemma notiert wurde. Zur Organisation eines einheitlichen Zugriffs auf derartige „Dubletten oder Mehrfachbildungen“ werden in Zukunft zusätzliche Informationen notwendig sein, z. B. durch Einführung von Meta- bzw. Mehrfachlemmata (vgl. REICHMANN 2012: 162 f.).

[62] Zur entsprechenden Ressourcenentwicklung gehört auch die kürzlich (März 2020) veröffentlichte erste Version eines Informationsservice (zunächst) zu niedersorbischen Ortsnamen: https://www.niedersorbisch.de/mjenja/. Die entsprechende Datenbasis soll sowohl um andere Namenstypen erweitert als auch perspektivisch auf das Obersorbische ausgedehnt werden.

Textsorte und -alter in der Regel Erkennungsquoten zwischen 93 und 96 Prozent erreicht. Das bedeutet: Zwischen vier und sieben Prozent der Tokens unterliegen zurzeit noch einer „Problembehandlung“.[63]

Für diese werden alle bis dahin aufgekommenen Fragen, Probleme und Komplikationen gesammelt und typisiert einer manuellen Nachbereitung zugeführt. Auf menügeführte Weise können so aufgetretene Probleme gelöst und Fehler sowohl in den KORPUSTEXTEN als auch in den Corproc-Programmteilen lokalisiert werden. Die gewählten Problemlösungen werden abstrahiert in die automatisierten Verarbeitungsschritte zurückgeführt, um diese zu vervollkommnen und so den manuellen Aufwand in nachfolgend zu bearbeitenden KORPUSTEXTEN zu verringern.

Im Rahmen der Problembehandlung sieht der Prozessflussplan (s. Kap. 4.5) einen aus Gründen der Qualitätssicherung dreistufigen Ablauf vor: die Problembehandlung selbst, nachfolgend die Evaluation des ersten Schritts und abschließend die tatsächliche Anwendung gewählter Problemlösungen, aufgeteilt in automatisiert und manuell umzusetzende Lösungen – letztere für Fälle, in denen eine Sichtung des XML-Quelltextes oder der analogen Vorlage des KORPUSTEXTES nötig ist. Außerdem werden Exporte an Mitarbeitende generiert, die außerhalb von Corproc tätig sind, sowie Feedbackrouten zwischen verschiedenen an der Bearbeitung beteiligten Rollen.

Unter den manuell zu bearbeitenden Fällen („Problemen“) sind folgende generell am häufigsten:

- `missLex`: bisher nicht registrierte Lexik einschließlich „neuer“ grammatischer Formen,
- `isName`: bisher nicht registrierte Eigennamen (zurzeit beschränkt auf bestimmte Typen),
- `isForeign`: zuvor nicht ausgeschlossene (vgl. Kap. 4.4.6) und bisher nicht registrierte „fremdsprachige“ Tokens.[64]

Außerdem werden Fehler bei der Tokenisierung[65] behoben – (ggf. fälschlich) zusammengeschriebene und daher zunächst als ein Token annotierte Formen können getrennt (`splitToken`), fälschlich getrennte zusammengeführt werden (`joinToken`). Ebenso werden mögliche Druck- oder Digitalisierungsfehler geklärt (`possTypo`). Dabei gilt stets das Primat der Originaltreue. Davon abweichende, weil nützliche Informationen werden auf der Annotationsebene ergänzt. Nur bei Digitalisierungsfehlern und wenigen anderen Fällen wird in den objektsprachlichen Teil des KORPUSTEXTES eingegriffen.

---

[63] In diesem Prozess waren als Corproc-Bearbeiter*in bzw. -Evaluator*in (im Sinne der in Kap. 4.5 genannten Rollen) beteiligt: Katja Atanasov, Thomas Menzel, Joanna Szczepańska sowie für das Obersorbische im Rahmen des Schrifttums-Monitorings (s. Kap. 4.6) Richard Bígl und seit 2020 auch Božena Braumanowa. Deren Rückmeldungen zu Programmfunktionen und Problemen bei der Entscheidungsfindung bzw. Annotation waren und sind von großer Bedeutung für die (Weiter-)Entwicklung der Software sowie der Annotationsrichtlinien (vgl. Fußnote 70).

[64] Einen häufigen Fall stellen die auch heute noch im „Nowy Casnik“ anzutreffenden deutschen Äquivalente für bisher (vermeintlich) noch nicht etablierte niedersorbische Wörter dar, die den sorbischen Begriffen meist in Klammern nachgestellt sind.

[65] Diese können wiederum Grundlage für eine Weiterentwicklung des Tokenisierers sein (s. Kap. 4.4.2).

In neueren Texten mit einem generell relativ hohen Grad an Normiertheit werden als besondere Formen „nicht-systemischer“ Lexik eingestreute dialektale oder archaische Textpassagen markiert und zunächst aus der weiteren Analyse ausgesondert (`isDialect, isOld`). In älteren Texten hat es sich mit Blick auf zahlreiche Schreibvarianten bewährt, bei an sich bekannten Wortformen sogleich eine entsprechende Zuordnung vorzunehmen (`editNorm`); dabei kann aus dem Vollformenlexikon ausgewählt werden (z. B. *łazechu* → *wlacechu* = 3. Person Plural Präteritum (Imperfekt) von *wlac* ‚schleppen‘ oder *Schleswig-Holftėinſkimi* → *schleswig-holsteinskimi*).

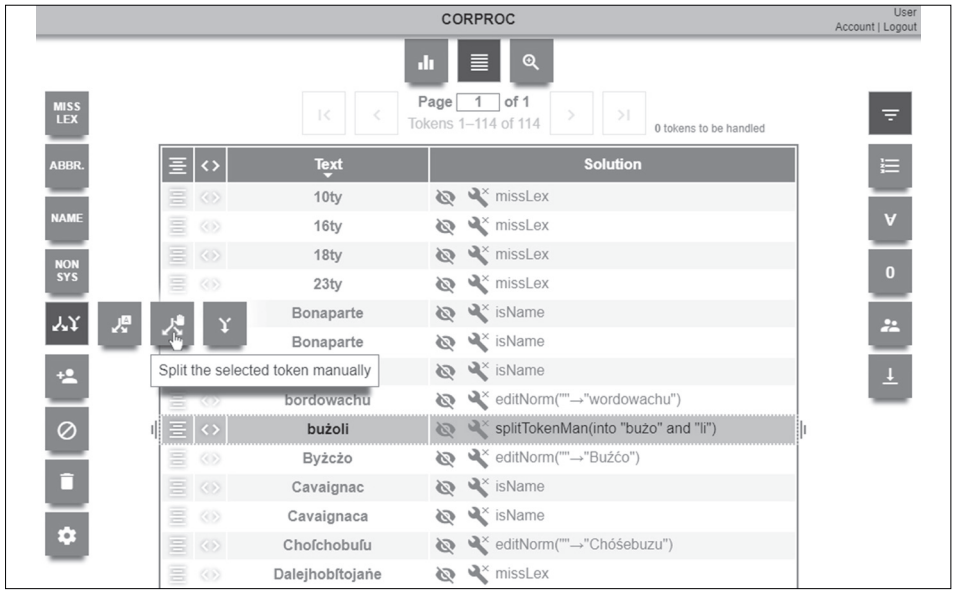


Abb. 5: Beispielansicht des Programms Corproc mit Bearbeitungsoptionen zur Problembehandlung

Derzeit findet diese Datennachbearbeitung ausschließlich auf Token-Ebene statt. Zur Bewertung des jeweiligen Falls liegen aber Kontextinformationen vor. Viele der oben genannten Kategorien sind im Einzelfall alles andere als klar abgrenzbar und die Bearbeiter*innen sind v. a. in älteren Texten mit einer Vielzahl unterschiedlicher Fälle konfrontiert. Eine entsprechende Problematisierung und Entscheidungshilfen bieten die Annotationsrichtlinien. Deren genauere Darstellung würde den Rahmen dieses Artikels sprengen.

Mit dem beschriebenen Vorgehen folgt auf die bisherigen Stufen der Qualitätsprüfung von Korpustexten damit hier eine dritte auf Token-Ebene.

4.4.6 Sprachkennzeichnung/Spracherkennung

Schon bei der Vorbereitung eines Korpustextes wird seine Hauptsprache – im Sinne der dominierenden Sprache unabhängig von abweichenden Passagen im Text – bestimmt und in den Metadaten festgehalten. Dies geschieht – über das Element `<language>` im `<teiHeader>` hinausgehend – im `xml:lang`-Attribut z. B. des Elements `<text>`. Für niedersorbische Texte ist die Kennzeichnung nach ISO-639-3 entsprechend `<text`

`xml:lang="dsb">`. Im Zuge einer detaillierteren textstrukturellen Annotation werden dann davon abweichende Passagen wiederum markiert.[66] In niedersorbischen KORPUSTEXTEN sind dies vor allem deutschsprachige Passagen. Dieser relativ aufwendige Prozess kann nur mit Blick auf gut erkennbare und längere Textfragmente erfolgen, mindestens für größere Wortgruppen. Diese können dann aber ggf. bei automatischen Verarbeitungsprozessen wie der Normalisierung oder der Lemmatisierung ausgeschlossen werden, wodurch zusätzliche Fehlerquellen vermieden werden können.[67] Das betrifft auch Tokens, die bei der Problembehandlung als „`isForeign`" markiert wurden.

Eine darüber hinausgehende, d. h. im Idealfall auch kleinere nicht-sorbische Wortgruppen oder Einzelwörter erfassende automatische Spracherkennung befindet sich noch in der Entwicklung.[68] Diese wird voraussichtlich der Tokenisierung nachgelagert sein und als probabilistisches Verfahren mithilfe von sprachspezifischen Präfixen, Suffixen und Buchstaben-n-Grammen umgesetzt.

Auch mit Blick auf Spracherkennung und -kennzeichnung spielen Eigennamen eine besondere Rolle, da natürlich auch in niedersorbischen Texten über Personen und Orte gesprochen wird, die nicht-sorbische Namen tragen (*Angela Merkel, Lissabon, Kuwait*). Daher ist die gesonderte Erfassung auch nicht-sorbischer Eigennamen von großer Bedeutung – die Grenzziehung ist überdies ohnehin problematisch.

4.4.7 Beschränkungen und Entwicklungsbedarf

Die Entwicklung des Gesamtverfahrens und seiner einzelnen Schritte sowie der dafür notwendigen Methoden und Instrumente ist inzwischen weit fortgeschritten, aber keineswegs abgeschlossen. Der Hauptgrund dafür ist, dass sich sämtliche Komponenten nun zunächst in größerem Umfang „am echten Textmaterial" bewähren müssen (vgl. Kap. 4.6). Außerdem gestaltet sich das Verfahren – das sollte deutlich geworden sein – insgesamt sehr aufwendig. Daher wurde die Lösung einiger bekannter Probleme der Korpusaufbereitung zunächst zurückgestellt.

Bereits erwähnt wurde die derzeit gültige Beschränkung auf eine Annotation auf Token-Ebene und damit der vorläufige Verzicht auf die Identifikation und Kennzeichnung von Mehr-Token-Einheiten (und damit von Mehrwortausdrücken). Außerdem werden bisher – neben unproblematischen, da eindeutig nicht-niedersorbischen Textpassagen – bestimmte „periphere" Textbestandteile zunächst aus der weiteren Analyse und Annotation ausgeschlossen (s. 4.4.5: `isOld, isDialect` sowie diverse symbolartige Tokens). Diese werden ggf. später einer Nachbearbeitung unterzogen, sind aber für den primär zu leistenden effektiven Textzugang tatsächlich nicht entscheidend.

---

[66] Da das Attribut `xml:lang` jeweils den Inhalt des eigenen Elements (z. B. eines Abschnitts `<p></p>`) sowie alle hierarchisch niedrigeren umfasst, müssen ggf. dort wiederum abweichende Passagen erneut gekennzeichnet werden, also etwa ein niedersorbischer Satz in einem ansonsten deutschsprachigen Absatz. In älteren Ausgaben des „Bramborski Casnik" betrifft dies vor allem Werbeannoncen: `<div type="advertisement" xml:lang="deu">`.

[67] Das deutsche Wort *Dom*, das in einem deutschsprachigen Absatz zu sakralen Gebäuden vorkommt, wird so von der sorbischen Lemmatisierung ausgeschlossen. Ansonsten würde es fälschlich als ns. *dom* ‚Haus' erkannt.

[68] Diese wurde in Grundzügen ebenfalls im ESF-Projekt „Sorbenwissen" (vgl. Fußnote 52) von Christopher Georgi als Mitarbeiter an der TU Dresden erarbeitet.

Ein grundsätzliches Problem bei der automatischen Normalisierung und Lemmatisierung (s. 4.4.3) ist die Möglichkeit von „false positives“, d. h. von „Treffern“, die nicht mehr zur „Problembehandlung“ (s. 4.4.5) vorgelegt werden, obwohl sich bei genauerer Prüfung des Einzelfalls zeigen würde, dass es sich um eine falsche Zuordnung durch die angewandten Algorithmen handelte. Die bisher implementierten Mechanismen versuchen dieses Phänomen klein zu halten, indem zum Beispiel Übergeneralisierungen bei der Normalisierung durch präzise Regel- oder Musterformulierungen möglichst vermieden werden. Auszuschließen sind „false positives“ aber nicht, und eine systematische Überprüfung der Normalisierungs- und Lemmatisierungszuordnungen in einer späteren Projektphase stellt einen weiteren wichtigen Schritt der Qualitätskontrolle dar.

Ähnlich sieht es mit der Auflösung von Homografen bei der Lemmatisierung aus: So sind bestimmte Textwörter ohne weitergehende Analyse (v. a. Bestimmung der Wortart und der syntaktischen Funktion) nicht nur einem Lemma zuzuordnen, sind also diesbezüglich mehrdeutig: Z. B. kann die Wortform/das Token *lěta* eine Flexionsform des Verbs *lětaś* ‚fliegen‘ (3. Person Singular Präsens), des Substantivs *lěto* ‚Jahr‘ (Genitiv Singular oder Nominativ/Akkusativ Plural) oder des Substantivs *lět* ‚Flug‘ (Genitiv Singular oder Nominativ/Akkusativ Dual) sein. Durch mögliche Schreibvarianten in älteren Texten wird die Mehrdeutigkeit noch größer. Eine spätere Disambiguierung ist notwendig. Darüber hinaus gibt es auch Mehrdeutigkeiten bei den zuzuordnenden Lemmata, z. B. *móc* als Verb (‚können‘) bzw. als Substantiv (‚Kraft, Macht, Gewalt‘). Hier enthält die lexikalische Datenbank schon mehr Information als derzeit bei der Lemmatisierung annotiert wird, sodass eine Nachbearbeitung zu gegebener Zeit erfolgen kann.

Es ließen sich weitere Beschränkungen nennen, die aber auf einer anderen Ebene liegen: So werden derzeit in den KORPUSTEXTEN noch keine Wortarten annotiert (part-of-speech-tagging). Eine derartige Annotation wäre vor allem für linguistische Auswertungen des Textkorpus von Bedeutung. Die zuvor erwähnten bereits entwickelten Sprachressourcen (morphologischer Analysator/Generator[69]) bieten für einen solchen Schritt jedoch eine gute Grundlage.

4.5 Gesamtworkflow[70]

Wie bereits in Kapitel 4.4.1 erwähnt, wurden für den Gesamtworkflow der Tätigkeiten, die durch das Programm Corproc organisiert werden oder die unmittelbar durch Datenaustausch und -bearbeitung aus/für Corproc beteiligt sind, insgesamt zehn Rollen unterschieden, die in wohldefinierten Beziehungen zueinander stehen. Diese Rollen sind:

1. Korpusverwalter*in: Verwaltung der KORPUSTEXTE für den Corproc-Input/Output; Dateivalidierung; Versionskontrolle

---

[69] Eine solche Datengrundlage existiert ebenfalls für das Obersorbische: https://www.soblex.de/.

[70] Der im Folgenden grob dargestellte Gesamtworkflow mit dem darin organisierten Zusammenspiel der beteiligten Rollen wurde über längere Zeit in der Projektgruppe entwickelt und aufgrund der praktischen Erfahrungen immer wieder modifiziert. Das Prozessflussdiagramm wurde auf Grundlage dieser Diskussionen und vorhergehender Versionen (Marcin Szczepański) durch Marek Slodička erstellt und fortlaufend weiterentwickelt. Die erwähnten Corproc-Annotations-Richtlinien wurden ebenfalls aus der Arbeitsgruppe heraus formuliert. Die Schriftfassung wurde von Thomas Menzel erstellt und wird von diesem verwaltet.

2. Projektverwalter*in: Organisation und Überwachung der Corproc-Abläufe, insbesondere der internen Dateiverwaltung (nach der Übernahme von 1. bis zur Rückgabe)
3. Corproc-Verwalter*in: u. a. Durchführung von Tokenisierung und Normalisierung/Lemmatisierung; Organisation der Problembehandlung; Datenexporte; Dateibereinigung um interne Metadaten vor Rückgabe an 1.; Einbindung neuer Versionen der Lexik-Datenbank und erneute Lemmatisierung
4. Corproc-Bearbeiter*in: Problembehandlung nach den Richtlinien
5. Corproc-Evaluator*in: Evaluation von Problemlösungsvorschlägen und Anwendung bzw. Klärung von Zweifelsfällen oder Einleitung weitergehender Klärungsprozesse
6. XML-Quelltext-Bearbeiter*in: Durchführung notwendiger direkter Eingriffe in den XML-KORPUSTEXT, ggf. unter Beachtung der analogen Vorlage bzw. des Bilddigitalisats
7. Corproc-Entwickler*in: Weiterentwicklung der Software, Behebung von Fehlfunktionen und Implementierung neuer Funktionen, Programmdokumentation
8. Lexik-Bearbeiter*in: Systematisierung und linguistische Bewertung der von 4./5. erarbeiteten Lösungen; Organisation der Weiterverarbeitung von Problem- und Zweifelsfällen
9. Lexik-Verwalter*in: Einarbeitung bisher in der lexikalischen Datenbank nicht erfasster Wortformen/Lexeme; Erstellung der neuen Lemmatisierungsdatenbank
10. Interne Sprachkommission: abschließende Lösung von Problemfällen; Änderung der Richtlinien; Entscheidung über notwendige Änderungen an Corproc

Das komplexe Zusammenspiel dieser Rollen ist in folgender Grafik[71] veranschaulicht:

[71] Der Abdruck der Grafik an dieser Stelle dient lediglich zur Veranschaulichung der Komplexität der Abläufe. Eine PDF-Fassung der Datei zur „Lektüre“ ist online verfügbar: https://niedersorbisch.de/download/gesamtworkflow-der-korpustext-bearbeitung-mit-corproc.pdf.

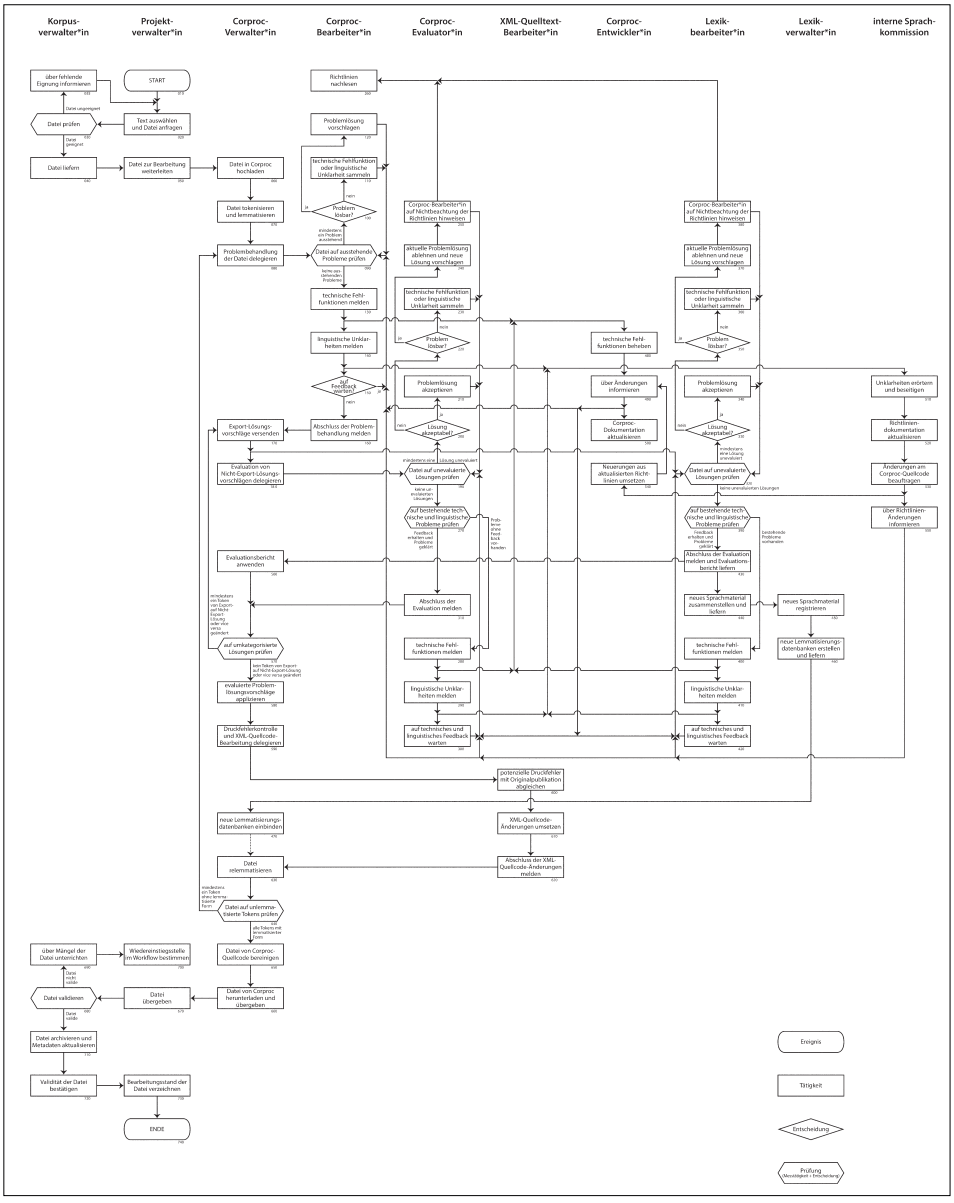


Abb. 6: Gesamtworkflow der KORPUSTEXT-Bearbeitung mit Corproc

4.6 Umsetzung: Referenzkorpora und Schrifttumsmonitoring

Das beschriebene Vorgehen wurde teilweise parallel zur Erstellung zweier sorbischer Referenzkorpora[72] entwickelt, sodass die dort eingehenden KORPUSTEXTE den ersten Anwendungs- und Testfall zentraler Teile des Verfahrens und damit eine wichtige Entwicklungsetappe darstellten. Die beiden Referenzkorpora[73] umfassen jeweils Text im Umfang von insgesamt etwa einer Million Tokens. Dabei handelt es sich um Textfragmente, die aus einer größeren, zuvor schon weitgehend nach den hier beschriebenen Richtlinien bearbeiteten Textbasis der Jahre 1990–2010 zufällig ausgewählt wurden. Für das Obersorbische wurde Ende 2018 eine automatische Lemmatisierungsquote von 93,5 % erreicht, für das Niedersorbische – vor allem aufgrund eines schon umfangreicheren Vollformenlexikons – von 96 %. Mitte 2020 wurden durch weiteren Ausbau des Lexikons für das Niedersorbische bereits Quoten von bis zu 98 % erreicht.

Seit Kurzem wird das inzwischen weiterentwickelte Verfahren nun beim laufenden Monitoring des nieder- und obersorbischen Schrifttums angewendet, ein Vorhaben, das seit 2019 in einer zweijährigen Pilotphase läuft.[74] Dazu erhält das SI auf Grundlage einer vertraglichen Regelung nahezu sämtliche im Domowina-Verlag (LND) erschienenen sorbischen Texte in bereits digitaler Form zur weiteren Bearbeitung. In deren Verlauf entstehen KORPUSTEXTE, die den beschriebenen Analyse- und Annotationsschritten unterworfen werden.[75] Ziel dieses korpuslinguistisch basierten Schrifttumsmonitorings ist eine laufende computergestützte „Beobachtung" der Sprachentwicklung zumindest im lexikalischen Bereich. Dabei sollen „neue" Wörter und Schreibungen erfasst und ggf. bewertet werden. Dabei kann es sich – neben (vermeintlichen) „Fehlern" unterschiedlicher Art – tatsächlich um Innovationen handeln, aber ebenso um ältere, jedoch bisher lexikografisch nicht erfasste Formen. Von Interesse wird auch sein, auf sich verändernde Verwendungshäufigkeiten zu achten, die für längerfristige Entwicklungen stehen können oder aber für „Konjunkturen" bei Themen und Diskursen.[76] Aufgrund der besonderen Sprachsituation ist nicht einmal ausgeschlossen, dass wir es auch im 21. Jahrhundert weiter mit stark personengebundenen Phänomenen zu tun haben, wie sie – am Beispiel des „Bramborski Casnik" – für das 19./20. Jahrhundert beschrieben wurden (BARTELS 2008). Die Ergebnisse des Monitorings werden u. a. eine empirische Datengrundlage für Wörterbücher liefern, z. B. hinsichtlich der Frage, welche Wörter neu aufgenommen werden sollten und

---

[72] Es handelte sich hierbei um ein Drittmittelprojekt, das von der Stiftung für das sorbische Volk gefördert wurde im Rahmen der vom Bund und den Ländern Sachsen und Brandenburg finanzierten Förderinitiative „Sorbisch in den neuen elektronischen Medien".

[73] Beide werden über die in Kap. 5.2.2 beschriebene neue „Komfortsuche" für Textkorpora zugänglich gemacht, zunächst für das Niedersorbische (Juni 2020). Dort findet sich auch eine genauere Beschreibung der Korpora.

[74] Auch dieses Projekt wird aus der in Fußnote 72 genannten Förderlinie finanziert.

[75] Die vom LND für 2019 gelieferten Texte hatten einen Umfang von insgesamt etwa 2,83 Millionen Tokens. Dabei betrug das Mengenverhältnis Obersorbisch:Niedersorbisch 2019 nahezu exakt 4:1: ca. 2,25 Millionen Tokens os. vs. 578 000 Tokens ns.

[76] Schlüsselwort-Analysen (Keyword Analysis) können bei entsprechender Datenbasis interessante Anstöße zu weiterer Forschung liefern. Dabei geht es um „auffällige" Gebrauchsfrequenzen bestimmter Wörter eines Textes bzw. einer Textmenge (Korpus) im Vergleich zu einem als „Standard" oder „Norm" definierten Textkorpus: „The notion of keyness is closely related to the notion of *aboutness*, that is, the understanding of the main concepts, topics or attitudes discussed in a text or corpus" (GABRIELETOS 2018: 225; Hervorhebung i. Orig.).

welche (in Druckfassungen) dafür gestrichen werden müssen. Ein erster öffentlicher Bericht ist nach Abschluss der Pilotphase und Auswertung der Daten für 2021 vorgesehen.

Ergänzend ist eine schrittweise Einbeziehung und Analyse älterer Texte in das Verfahren vorgesehen, gewissermaßen als „retrogrades Schrifttumsmonitoring“. Langfristiges Ziel ist eine möglichst vollständige lexikalische Analyse des niedersorbischen Schrifttums (vgl. Kap. 4.1). Einen Auftakt hierfür stellen die Arbeiten zur Vorbereitung einer niedersorbischen digitalen Bibliothek dar (vgl. Kap. 5.2.3).

5. Digitalisierungskette: Fokussierungen und Zugänge

Betrachtet man KORPUSTEXTE in einem ganzheitlichen Ansatz als moderne und möglichst tief erschlossene computerlesbare Repräsentationen einzelner Instanzen des Schrifttums, so ist mit Blick auf sinnvolle Zugänge zu den entsprechenden Daten die gesamte Digitalisierungskette zu betrachten:

```
Objekt  >  Bilddigitalisat  >  ROHTEXT  >  KORPUSTEXT
                               [---------- Volltext ----------]
```

Grundsätzlich sind Zugänge auf allen Stufen dieser – hier ohne Zwischenstufen vereinfacht dargestellten – Digitalisierungskette möglich. Dabei spielen bei der Schaffung von Zugängen zu den (Teil-)Ergebnissen der Digitalisierung offensichtlich Fokussierungen eine Rolle.

5.1 Fokussierungen

5.1.1 Fokus Objekt-Bilddigitalisat

Die Digitalisierung von Schrifttum erfolgte meines Wissens bisher einerseits ausgehend vom und mit Fokus auf das physische Objekt. Dieses wird in obiger Darstellung der Digitalisierungskette sowie im Folgenden kurz als *Objekt* bezeichnet. Beispielsweise wird ein Buch durch Scannen zu einem (möglichst hochwertigen) Bilddigitalisat. Optional erfolgt eine automatische Texterkennung per OCR. Dies ist der Ansatz von Bibliotheken und entsprechend ausgerichteten Digitalisierungsprogrammen. Das Hauptaugenmerk liegt hier auf einer möglichst originalgetreuen und hochauflösenden Image-Digitalisierung. Diese Digitalisate als das Original vertretende Grafik-Bündel sind im Idealfall weltweit zugänglich und machen einen Zugriff auf das physische Original in den allermeisten Anwendungsfällen verzichtbar.

Sofern hierbei Volltexte erstellt werden, wie in den letzten Jahren für Digitalisierungsmaßnahmen zunehmend gefordert,[77] sind diese je nach Vorlagenqualität, Sprache(n), Schrifttype(n), Druck- bzw. Papierqualität, OCR-Software, Aufwand beim Softwareeinsatz und einer möglichen Nachbearbeitung usw. mehr oder weniger brauchbar. In man-

[77] Siehe DFG-Praxisregeln „Digitalisierung“ (DFG-Vordruck 12.151, Stand 12/2016), S. 8.

chen Fällen gestatten sie eine effektive Volltextsuche. Bezüglich der Qualität der technisch erreichbaren Volltexte ist der Ansatz hierbei pragmatisch, was sowohl mit Blick auf den Fokus der Relation Objekt-Bilddigitalisat als auch die Grenzen des technisch und finanziell bzw. personell Machbaren verständlich ist. Hauptziel ist jedenfalls in aller Regel nicht, um jeden Preis möglichst hochwertige Volltexte zu erzeugen – de facto setzt das im jeweiligen Fall per OCR Mögliche die Grenzen des Erreichbaren. In diesem Bereich gibt es aber derzeit eine starke Entwicklung zu mehr Leistungsfähigkeit (vgl. die Fußnoten 32 und 46).

Zugänge zu den Produkten derartiger Digitalisierungsmaßnahmen bedienen den linken Teil der oben dargestellten Digitalisierungskette: Objekt > Bilddigitalisat > (Volltext). Die Gesamtkette reißt hier entweder im oder vor dem Bereich der Volltexte ab.

5.1.2 Fokus Volltext

Einen anderen Ansatz verfolgen Textdigitalisierungsmaßnahmen, die eine korpuslinguistische Nutzung von Schrifttums-Instanzen anstreben. Auch hier lassen sich mehr oder weniger pragmatische Vorgehensweisen unterscheiden, je nachdem, wie viel Wert auf die Qualität der Volltexte gelegt wird. Big-Data-Methoden sind diesbezüglich genügsamer als der am Sorbischen Institut verfolgte, durch die „Small-Data-Problematik" begründete Weg. Auf jeden Fall steht hier die Erstellung (je nach Ziel: angemessen) nutzbarer Volltexte im Fokus; aus SI-Perspektive sogar noch weitgehender die Erstellung hochwertiger KORPUSTEXTE. Auch hier dienen so gut wie immer Bilddigitalisate als Vorlage für die Volltexterstellung. Sofern noch nicht vorhanden, müssen zu ihrer Herstellung auch die Objekte herangezogen werden. Aber zumindest in früheren Phasen war es die Relation Objekt-Bilddigitalisat, auf die mit großem Pragmatismus geschaut wurde: Die Bilddigitalisate mussten allein den Anforderungen der Volltexterstellung genügen. Die heute anerkannten allgemeinen Vorschriften einer nachhaltigen Bilddigitalisierung[78] wurden häufig nicht beachtet.

In diesem Ansatz wird der rechte Teil der Digitalisierungs-Gesamtkette bedient: KORPUSTEXT < ROHTEXT < (Bilddigitalisat). Die Kette reißt hier im oder (zumindest mit Blick auf heute geltende Standards) vor dem Bereich der Bilddigitalisate ab.

Ziel der am Sorbischen Institut verfolgten Strategie ist es, die beiden oben skizzierten Ansätze bzw. Fokussierungen in einem ganzheitlichen Ansatz der Schrifttumsdigitalisierung miteinander zu verbinden.[79]

---

[78] DFG-Praxisregeln „Digitalisierung" (DFG-Vordruck 12.151, Stand 12/2016), S. 13 ff.

[79] In eine solche Richtung ist auch der Beitrag von BUBENHOFER/ROTHENHÄUSLER (2016) zu deuten, wo mit Blick auf die Bibliotheken das Fazit gezogen wird: „Die Bibliotheken müssen ihre Bestände im Volltext, nicht nur deren Metadaten digital verfügbar machen, so dass mit computergestützten Verfahren darauf Forschung betrieben werden kann. […] Sie wären also nicht nur Bibliothek, sondern auch ‚Korporathek'" (S. 69). Es bleiben freilich die in diesem Artikel thematisierten Fragen der Qualität der Volltexte sowie die Art ihrer Aufbereitung. Letztlich ist nicht entscheidend, ob eine angemessene Lösung primär aus der Perspektive der KORPUSTEXTE oder einer bibliothekarischen Primärdigitalisierung gedacht wird. Entscheidend ist, dass überhaupt eine Gesamtperspektive eingenommen wird und die Maßnahmen aller Stufen koordiniert sind.

5.2 Zugänge

Dem thematischen Zuschnitt dieses Artikels entsprechend geht es im Folgenden um bisherige, kürzlich realisierte und geplante öffentliche Datenzugänge auf Volltexte. Nur der Vollständigkeit halber sei erwähnt, dass der Zugang zu den sorbischen „Objekten" natürlich auf übliche Weise gewährleistet ist: über einen (in)direkten Zugriff auf Präsenz- und Magazinbestand in den Räumlichkeiten der Sorbischen Zentralbibliothek (SZB) am Sorbischen Institut[80], ggf. mit vorgelagerter Metadaten-Recherche in den Bibliotheks- und Verbundkatalogen. Die von der SZB bzw. dem Kooperationspartner, der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB) bereitgestellten Bilddigitalisate sind über das Webportal Sachsen.digital[81] zugänglich.

5.2.1 Rohtext

Seit Ende 2010 ist über das Sprachportal niedersorbisch.de ein Zugang zu niedersorbischen Volltexten möglich.[82] Auf der Internetseite ist von einer „Sammlung niedersorbischer Texte" die Rede, und tatsächlich handelt es sich dabei nach der in diesem Artikel verwendeten Terminologie um Rohtexte im Umfang von etwa 15 Millionen Tokens. Aus rechtlichen Gründen ist dieser Zugang auf bis 1937 erschienene Texte beschränkt; eine Quellenliste findet sich auf der Internetseite. Dieses „niedersorbische Textkorpus" ist der öffentlich zugängliche Teil des „alten" Textkorpus (vgl. Kap. 1).

Dieser erste und weiterhin aktive Zugang – nach der im Folgeabschnitt dargestellten Erweiterung ist dies die Unterseite „Standardsuche im alten Korpus" – wurde ermöglicht durch technische Unterstützung des Ústav Českého národního korpusu der Prager Karls-Universität. Von dessen Seiten aus kann ebenfalls auf die niedersorbischen Texte des alten Korpus zugegriffen werden.[83] Fortgeschrittenen Nutzerinnen und Nutzern stellt der Prager Zugang zusätzliche Funktionen (Filter, Kollokationsabfrage usw.) zur Verfügung, die über das Interface auf niedersorbisch.de nicht möglich sind. (Außerdem ist hier auch das obersorbische Textkorpus zugänglich.[84] Es hat mit Blick auf die Unterscheidung von Rohtext und Korpustext bzw. „alt" vs. „neu" denselben Status wie das in diesem Kapitel thematisierte niedersorbische Korpus.)

Da es sich um Rohtexte handelt, ist die Qualität sehr unterschiedlich, und die Texte sind kaum bearbeitet, d. h. auch weder normalisiert noch lemmatisiert. Die Suche ist entsprechend anspruchsvoll. Informationen zu historischen Schreibkonventionen sowie zur

---

80 Internet: https://www.serbski-institut.de/de/Bibliothek-und-Archiv/ [15.07.2020].

81 Internet: https://sachsen.digital/. Geplant ist auch eine Präsentation auf dem sorabistischen Wissensportal SORABICON (http://www.sorabicon.de) [15.07.2020].

82 Internet: https://www.niedersorbisch.de/korpus/; der oben thematisierte und vormals auf dieser Seite einzige Zugang findet sich, nach der Erweiterung der Seite um eine „Komfortsuche" in Korpustexten (vgl. Kap. 5.2.2) nun unter der Adresse https://www.niedersorbisch.de/korpus/standard/ [15.07.2020].

83 Internet: https://kontext.korpus.cz (Auswahlkorpus: „dotko" – Abkürzung für Dolnoserbski tekstowy korpus) oder https://kontext.korpus.cz/first_form?corpname=dotko [05.05.2020].

84 Wie in vorangehender Fußnote, nur „hotko" (Hornjoserbski tekstowy korpus) statt „dotko". Zum obersorbischen Korpus s. Wölkowa 2013, 2014 sowie Kaulfürst 2014b.

Nutzung von Regulären Ausdrücken versuchen, dieses Problem abzumildern.[85] Gleichwohl sind die Zugangsbarrieren relativ hoch (vgl. Kap. 3).

Wie der folgende Screenshot zeigt, bietet der Zugang die Funde in üblicher KWIC[86]-Ansicht mit dem Suchwort in einem Standard-Kontext von acht Tokens links wie rechts. Der einzelne Kontext kann hier auf eine Größe von insg. 300 Tokens erweitert werden.[87] Vollständige Texte sind nicht zugänglich.

Abb. 7: KWIC-Ansicht zu ROHTEXTEN über niedersorbisch.de

5.2.2 Korpustext

Ein Zugang zu KORPUSTEXTEN und damit auf den ersten Teil des „neuen" niedersorbischen Textkorpus ist erst seit Kurzem online: Im Juni 2020 wurde auf niedersorbisch.de eine „Komfortsuche" veröffentlicht, die erstmals eine Recherche in den qualitätsgesicherten und lemmatisierten Volltexten ermöglicht.[88]

Dieser neue Zugang unterscheidet sich vom bisherigen zunächst durch eine andere Textbasis: Statt älterer Rohtexte unterschiedlicher Qualität werden hier nach und nach ausschließlich hochwertig digitalisierte KORPUSTEXTE bereitgestellt. Durch die zusätz-

---

85 Die Seite wurde von Marcin Szczepański erstellt, die Rahmentexte und Hilfestellungen von Fabian Kaulfürst.

86 Key-Word-In-Context, d. h. das Suchwort in der Mitte von Kontext links und rechts umrahmt.

87 Über den oben erwähnten Prager Zugang sind größere Kontexte möglich.

88 Dieses neue Interface zur Korpussuche wurde in der zweiten Hälfte 2019 von Marek Slodička implementiert und in Zusammenarbeit mit Marcin Szczepański in das Sprachportal niedersorbisch.de integriert. Es basiert auf der am Institut für maschinelle Sprachverarbeitung der Universität Stuttgart entwickelten Corpus Workbench (EVERT/HARDIE 2011). Das Projekt wurde gefördert vom Ministerium für Wissenschaft, Forschung und Kultur des Landes Brandenburg. Internet: https://niedersorbisch.de/korpus/, Unterseite „Komfortsuche im neuen Korpus" oder direkt: https://niedersorbisch.de/korpus/komfort/pytanje/ [15.07.2020].

lich erfolgte Aufbereitung (v. a. Normalisierung und Lemmatisierung) ist es möglich, nach niedersorbischen „Wörtern" (im Sinne von Lemmata, wie sie in Wörterbüchern zu finden sind) zu suchen, wobei alle (Schreibungen von) Flexionsformen und damit die Gesamtheit aller Belege eines Wortes angezeigt werden. Damit präsentiert dieser Zugang nicht nur höherwertige und besser aufbereitete KORPUSTEXTE, sondern senkt auch in erheblichem Maße die in Kap. 3 thematisierten Nutzungsbarrieren. Denn eine Suche nach der niedersorbischen Standard-Form *cas* ‚Zeit' findet auch die Wortformen bzw. Schreibvarianten *casom, zaß, zaffoju* usw. Auch die Suche nach Kollokationen ist durch einfache Kombination von – durch Leerzeichen getrennten – Suchwörtern auf diese Weise möglich: so ergibt eine Suche nach *wón byś* ‚er sein' auch Ergebnisse wie *jomu było* ‚ihm war'. Eine „Expertensuche" bietet auf Wunsch weitere Optionen. Eine detaillierte Beschreibung findet sich auf der Internetseite. Beschränkungen bei der Suche ergeben sich vor allem aus dem derzeitigen Stand der Aufbereitung der KORPUSTEXTE (s. Kap. 4.4.7). Damit sind in Zukunft weitere Verbesserungen zu erwarten.

Die jeweiligen Anzeigemöglichkeiten werden sich in Abhängigkeit vom Charakter der Textbasis (z. B. von „normalen" KORPUSTEXTEN im Unterschied zum aus Textfragmenten bestehenden Referenzkorpus; vgl. Kap. 4.6) unterscheiden. Ein vollständiger Zugriff auf die Texte muss rechtlich möglich sein. Dies ist derzeit vor allem bei Gemeinfreiheit der Fall. Auch der Funktionsumfang des neuen Zugangs soll zukünftig erweitert werden.

5.2.3 KORPUSTEXT-basierte Digitale Bibliothek

Der Begriff *Digitale Bibliothek* meint hier eine öffentliche Bereitstellung vollständiger und in Grundzügen textstrukturell (nach)modellierter digitaler Lesefassungen von (zuvor retrodigitalisierten) Instanzen des sorbischen Schrifttums.[89] Im Unterschied zu den in der Korpuslinguistik üblichen KWIC-Ansichten unterschiedlichen Umfangs (s. Abb. 7) stehen damit hier grundsätzlich die kompletten Texte zur Verfügung. Die textstrukturelle Annotation der zugrundeliegenden KORPUSTEXTE (s. 4.3.3) ermöglicht außerdem eine Darstellung, die sich von reinen Fließtexten, die man gelegentlich aus KWIC-Ansichten heraus ansteuern kann, unterscheidet und sich traditionellen Druckdarstellungen annähert (s. u.).[90] Die im vorherigen Kapitel erwähnten komfortablen Suchmöglichkeiten auf Basis der zuvor erfolgten Normalisierung und Lemmatisierung der KORPUSTEXTE sollen auch hier möglichst weitgehend nutzbar sein.[91]

---

[89] Damit wird die Bezeichnung *Digitale Bibliothek* hier in einer anderen Bedeutung verwendet als in Bezug auf (mehr und mehr) digitale Angebote oder „digitalisierte" Teilbereiche klassischer Bibliotheken. In letzterem Sinne ist zum Beispiel häufig die Rede davon, dass eine entsprechende Einrichtung nun schrittweise in eine „digitale Bibliothek" überführt wird.

[90] Eine vollständig originalgetreue Modellierung der formalen Textstruktur, z. B. zur korrekten Darstellung dort vorhandener Wort- oder Zeilentrennungen, ist jedoch nicht Ziel unserer textstrukturellen Annotation. Siehe auch Fußnote 95.

[91] Offen ist noch, wie die in 5.2.2 beschriebene KORPUSTEXT-Nutzerschnittstelle systematisch mit der Infrastruktur der Digitalen Bibliothek verknüpft wird.

Bei den über diesen Zugang einer niedersorbischen[92] Digitalen Bibliothek bereitzustellenden Texten handelt es sich ausdrücklich nicht um kritische und den entsprechenden wissenschaftlichen Ansprüchen genügende Editionen (s. Jannidis/Kohle/Rehbein 2017: 237 ff.).[93] Es geht vielmehr – bezüglich möglicher Zugänge – um e i n Element des Versuchs, einen „Lückenschluss" zwischen den beiden oben dargestellten Fokussierungen der Digitalisierungskette herzustellen. Außerdem soll dem Wunsch entsprochen werden, hochwertige KORPUSTEXTE – über KWIC-Ansichten usw. hinaus – auch in gut handhabbaren Lesefassungen anzubieten, in denen wichtige Eigenschaften effektiv nutzbarer (Korpus)Texte miteinander kombiniert sind:

1. eine hohe Digitalisierungsgenauigkeit, allerdings ohne a b s o l u t e Fehlerfreiheit garantieren zu können;
2. eine an das Original angelehnte, aber nicht vollständige (Nach-)Modellierung der ursprünglichen Textgestalt durch textstrukturelle Annotation, wodurch eine dem Drucktext ähnliche Gestalt und Lesbarkeit erreicht werden kann;
3. eine Durchsuchbarkeit des Textes, die sich über die Textwort/Token-Ebene hinaus auf weitergehende Annotationen stützen kann, sodass der Zugang über normalisierte Lemmata (Grundformen der Textwörter) möglich ist;
4. die Ermöglichung einer gezielten Suche nach Eigennamen;[94]
5. bei Originalen in Frakturschrift die Option, sich die Texte auch in lateinischer Transliteration anzeigen zu lassen;
6. im Fall ausgewählter Texte die Möglichkeit, neben der originalen Schreibweise (in Fraktur oder Latein) auch eine normalisierte, d. h. orthografisch modernisierte Fassung zu lesen.

Da in dem entwickelten Workflow zur Erstellung hochwertiger KORPUSTEXTE die für eine digitale Textpräsentation im obigen Sinne notwendigen Bearbeitungsschritte (u. a. textstrukturelle Annotation, Normalisierung und Lemmatisierung) stets durchlaufen werden, können nach Etablierung der notwendigen Technik und Infrastruktur alle entsprechend aufbereiteten Texte in die KORPUSTEXT-basierte Digitale Bibliothek einfließen, sofern eine vollständige Textpräsentation rechtlich zulässig ist. Außerdem lässt sich die zu einem bestimmten Zeitpunkt erreichte Qualitätsstufe der Normalisierung, Lemmatisierung und Eigennamenkennzeichnung im Laufe der Zeit für einzelne Texte bzw. Textgruppen weiter steigern, indem sie die weitgehend automatisierten Analyse- und Annotationspro-

---

[92] Das Konzept einer KORPUSTEXT-basierten Digitalen Bibliothek ist auf Grundlage des dargestellten Entwicklungsrahmens auf beide sorbische Sprachen anwendbar. Es wird aber zunächst für das Niedersorbische umgesetzt, und zwar in einem Teilprojekt des Drittmittelvorhabens „Inwertsetzung des immateriellen Kulturerbes im deutsch-slawischen Kontext". Dabei handelt es sich um ein als Sofortmaßnahme im laufenden Strukturwandel in der Lausitz deklariertes Projekt, das 2019/20 vom Beauftragten der Bundesregierung für Kultur und Medien über das Ministerium für Wissenschaft, Forschung und Kultur des Landes Brandenburg gefördert wird.

[93] Solche könnten aber auf der Basis schon gut aufbereiteter KORPUSTEXTE der Digitalen Bibliothek entstehen.

[94] Diese ist derzeit noch stark eingeschränkt, wird aber in den nächsten Jahren durch fortschreitende Eigennamen-Erkennung und -Kennzeichnung in den KORPUSTEXTEN laufend erweitert (vgl. Kap. 4.4.4).

zesse mit einer angereicherten lexikalischen Datenbank erneut durchlaufen. Die Zuverlässigkeit der Textsuche z. B. nach Eigennamen wird dadurch perspektivisch schrittweise in Richtung 100-prozentiger Genauigkeit erhöht.

Das beschriebene Modell der Texterschließung stellt einen Kompromiss zwischen Erschließungsbreite und -tiefe dar und versucht einen angemessenen Aufwand für Textanalyse und -annotation bei Ermöglichung eines größtmöglichen Nutzens zu wahren. Dies bedeutet mit Blick auf die textstrukturelle Annotation, dass auf eine aufwendige originalgetreue Modellierung zum Beispiel von Tabellen und Listen verzichtet wird.[95] Die auf diese Weise und mit diesen Beschränkungen bearbeiteten Texte können aber jederzeit als Grundlage für einer tiefere bzw. detailgetreuere Modellierung verwendet werden. Eine erste Fassung der KORPUSTEXT-Bibliothek wird voraussichtlich zur Jahreswende 2021/22 online gehen.

In gewissem Sinne stellt die Mitte 2018 veröffentlichte Online-Fassung der ersten und bis heute einzigen niedersorbischen Bibel-Gesamtausgabe von 1868 einen Prototyp für die KORPUSTEXT-basierte Digitale Bibliothek dar.[96] Wichtige Eigenschaften dieser zum 150. Jubiläum der Bibelausgabe vorbereiteten Internet-Fassung sind:

- Möglich ist die Suche in originaler wie auch normalisierter (heutiger) Orthografie – Letzteres ist die Standard-Einstellung. Im Unterschied zu den KORPUSTEXTEN der Digitalen Bibliothek ist der Bibel-Text bisher nicht lemmatisiert, sodass die Textsuche tokenbasiert (auf Textwörter zugreifend) ist.
- Neben dieser Volltextsuche ist der Text über die für die Bibel übliche Textstruktur (Bücher u. a., Kapitel, Verse) zugänglich. Ein „Blättern“ zum jeweils vorangehenden bzw. nachfolgenden Abschnitt ist möglich.
- Der Text wird in drei Varianten angeboten: Grundlage für die ersten beiden – entweder in originaler Frakturschrift[97] oder in lateinischer Transliteration – ist eine genaue Abschrift der gedruckten Version in originaler Schreibung. Die dritte Version präsentiert den Originaltext in heutiger Orthografie.[98]

---

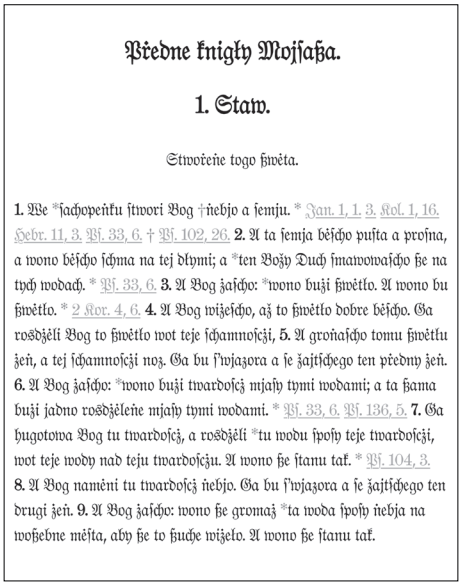
[95] Derzeit sind aus einer vollständigen textstrukturellen Modellierung ausgenommen: Tabellen, Listen, Gedichte sowie Anzeigen/Annoncen. Fotos u. a. Grafiken werden generell nicht angezeigt – stattdessen findet sich ein Platzhalter. (Ggf. vorhandene Bildunterschriften gelten hingegen als normaler Text.) Es ist möglich, dass einige dieser Elemente später nachmodelliert werden oder dass zumindest für bestimmte Textsorten eine komplette Modellierung erfolgt. Da mit der digitalen Darstellung das Faksimile der zugrundeliegenden Druckausgabe verknüpft wird, können diese Elemente bei Bedarf dort eingesehen werden.

[96] Internet: https://www.niedersorbisch.de/biblija/ [15.07.2020]. Die Online-Ausgabe basiert auf einem bereits 2005 im Double-Keying-Verfahren digitalisierten ROHTEXT. Dieser wurde für die Online-Veröffentlichung weiter aufbereitet, jedoch nicht vollständig nach dem in diesem Artikel beschriebenen Verfahren. Dafür fand in Kooperation mit dem Verein zur Förderung der wendischen Sprache in der Kirche e. V. eine komplette Korrekturlesung des Textes statt, sodass die Bibel-Ausgabe einen Schritt weiter in Richtung „Edition“ geht. Dies wird wegen des hohen Aufwandes höchstens bei ausgewählten Texten der KORPUSTEXT-basierten Digitalen Bibliothek möglich sein. – Weitergehende Informationen finden sich auf der Internetseite. Die Ausgabe wurde von Fabian Kaulfürst und Marcin Szczepański besorgt.

[97] Hierfür wurde von Marcin Szczepański der originale Bibel-Font digitalisiert: Es handelt sich dabei um eine exakte Retrodigitalisierung des Zeichensatzes, der seinerzeit beim Druck der sorbischen Bibel genutzt wurde. Dieser Unicode-Font beinhaltet sämtliche im Bibeltext aus dem Jahr 1868 auftretenden Buchstaben.

[98] Diese wurde von Fabian Kaulfürst durch eine aufwendige halbautomatische Transformation erstellt.

– Es handelt sich um einen Hypertext auch insofern, als die zahlreichen Verweise im Bibeltext als Links funktionieren.



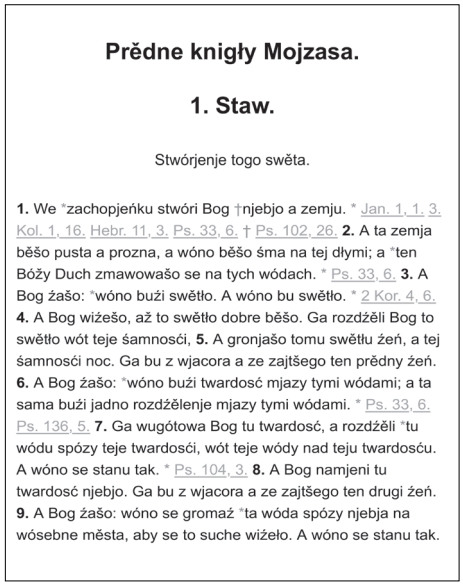
Pſedne knigły Mojſaßa.

1. Staw.

Stwořeńe togo ſwěta.

1. We *ſachopeńku ſtwori Bog †ńebjo a ſemju. * Jan. 1, 1. 3. Kol. 1, 16. Hebr. 11, 3. Pſ. 33, 6. † Pſ. 102, 26. 2. A ta ſemja běſcho puſta a proſna, a wono běſcho ſchma na tej dłymi; a *ten Boży Duch ſmawowaſcho ſe na tych wodach. * Pſ. 33, 6. 3. A Bog żaſcho: *wono buźi ſwětło. A wono bu ſwětło. * 2 Kor. 4, 6. 4. A Bog wiźeſcho, aż to ſwětło dobre běſcho. Ga rosdźěli Bog to ſwětło wot teje ſchamnoſcźi, 5. A groňaſcho tomu ſwětłu żeń, a tej ſchamnoſcźi noż. Ga bu ſ'wjazora a ſe żajtſchego ten předny żeń. 6. A Bog żaſcho: *wono buźi twardoſcź mjaſy tymi wodami; a ta ſama buźi jadno rosdźěleńe mjaſy tymi wodami. * Pſ. 33, 6. Pſ. 136, 5. 7. Ga hugotowa Bog tu twardoſcź, a rosdźěli *tu wodu ſpoſy teje twardoſcźi, wot teje wody nad teju twardoſcźu. A wono ſe ſtanu tak. * Pſ. 104, 3. 8. A Bog namėni tu twardoſcź ńebjo. Ga bu ſ'wjazora a ſe żajtſchego ten drugi żeń. 9. A Bog żaſcho: wono ſe gromaż *ta woda ſpoſy ńebja na woſebne měſta, aby ſe to ſuche wiźeło. A wono ſe ſtanu tak.

Ansicht in Originalschreibung (Fraktur)



Prědne knigły Mojzasa.

1. Staw.

Stwórjenje togo swěta.

1. We *zachopjeńku stwóri Bog †njebjo a zemju. * Jan. 1, 1. 3. Kol. 1, 16. Hebr. 11, 3. Ps. 33, 6. † Ps. 102, 26. 2. A ta zemja běšo pusta a prozna, a wóno běšo śma na tej dłymi; a *ten Bóży Duch zmawowašo se na tych wódach. * Ps. 33, 6. 3. A Bog žašo: *wóno buźi swětło. A wóno bu swětło. * 2 Kor. 4, 6. 4. A Bog wiźešo, až to swětło dobre běšo. Ga rozdźěli Bog to swětło wót teje śamnosći, 5. A gronjašo tomu swětłu žeń, a tej śamnosći noc. Ga bu z wjacora a ze zajtšego ten prědny žeń. 6. A Bog žašo: *wóno buźi twardosć mjazy tymi wódami; a ta sama buźi jadno rozdźělenje mjazy tymi wódami. * Ps. 33, 6. Ps. 136, 5. 7. Ga wugótowa Bog tu twardosć, a rozdźěli *tu wódu spózy teje twardosći, wót teje wódy nad teju twardosću. A wóno se stanu tak. * Ps. 104, 3. 8. A Bog namjeni tu twardosć njebjo. Ga bu z wjacora a ze zajtšego ten drugi žeń. 9. A Bog žašo: wóno se gromaź *ta wóda spózy njebja na wósebne města, aby se to suche wiźeło. A wóno se stanu tak.

Ansicht in moderner Orthografie

Abb. 8: Zwei der drei möglichen Ansichten des Bibeltextes

Trotz der Abweichungen im Detail ist diese Bibelausgabe das erste Beispiel für eine KORPUSTEXT-basierte digitale „Edition“. Ausgehend von der Annahme, dass das beschriebene Verfahren der Textdigitalisierung und -erschließung eine relativ hohe und für die formulierten Ziele hinreichende Textqualität liefert, stellt diese Art der Präsentation von gemeinfreien Texten einen zusätzlichen Zugang neben den üblichen KWIC-Ansichten dar. Sie ist vor allem für Nutzer des Textkorpus geeignet, die an größeren Kontexten von Fundstellen interessiert sind oder bestimmte Texte generell in Gänze lesen oder gezielt auswerten möchten. Dabei ermöglicht die Tatsache, dass die Ansichten auf annotierten Volltexten basieren, zugleich deren effektive Durchsuchbarkeit.

6. Fazit

Der in diesem Artikel beschriebene Ansatz zum Aufbau einer auf hochwertigen KORPUSTEXTEN gründenden digitalen Repräsentation des niedersorbischen Schrifttums ist ganzheitlich in dem Sinne, dass er verschiedene Aspekte und Interessen, die sich um digitale Texte gruppieren, systematisch miteinander zu verbinden sucht: Das im Aufbau befindliche historische Textkorpus (Vollkorpus) der niedersorbischen Schriftsprache und dabei insbesondere das GLOBALKORPUS (als möglichst große Teilmenge davon) bestehen aus KORPUSTEXTEN, die nicht nur eine valide Datengrundlage für die Forschung bereitstellen, sondern zugleich der Sprachdokumentation dienen. Ein im Aufbereitungsprozess verankerter mehrstufiger Prozess der Qualitätssicherung schafft hochwertige und damit

zuverlässige KORPUSTEXTE. Die dargestellte aufwendige Basis-Aufbereitung für das GLOBALKORPUS ermöglicht deren effiziente und barrierearme Auswertung nicht nur für linguistische, sondern auch für kulturwissenschaftliche und andere Fragestellungen. Dies wird die textbasierte Forschung zum Sorbischen hoffentlich befördern.

Gleichzeitig wird mit dem geschilderten Aufbereitungsverfahren, das unter spezifischen Umständen eine Balance zwischen automatisierten und manuellen Verfahren versucht, zugleich der Weg bereitet für einen möglichst unbeschränkten und komfortablen Zugang zu den Einzeltexten und damit zum niedersorbischen Schrifttum – bis hin zur Bereitstellung in einer Digitalen Bibliothek. So wird ein wichtiger Teil des niedersorbischen Kulturerbes auch als „Wissensspeicher“[99] bereitgestellt, als recherchierbares Abbild einer sich zum Beispiel in den historischen Zeitungen spiegelnden kulturellen Praxis.

Die auf diese Weise am Sorbischen Institut betriebene Öffnung (korpus)linguistischer Konzepte für auch kulturwissenschaftliche Forschung und die Bereitstellung einer entsprechenden Datengrundlage entspricht sich abzeichnenden Entwicklungen in diesem Bereich. So heißt es in einem neueren Sammelband zur Korpuslinguistik: „Es wäre zu wünschen, dass unter dem interdisziplinären Dach der Digital Humanities die digitale Aufbereitung von Sprachdaten in Zukunft auf eine Weise erfolgt, die Anschlussmöglichkeiten in alle relevanten fachlichen Richtungen offenhält. […] Die enge Definition des linguistischen Korpus, als digitale Textsammlung, die von Sprachwissenschaftlern für die Zwecke sprachwissenschaftlicher Analyse erstellt wurde, gilt nur mehr bedingt.“ (MAIR 2018: 11, 23)

Die bereits zu Beginn des Artikels betonte große strategische Bedeutung hochwertiger Textkorpora für die Sorabistik ist Motivation und Begründung für das hier vorgestellte Konzept, bei dessen Umsetzung in den letzten Jahren erhebliche Fortschritte erzielt wurden. Die Arbeit zum Niedersorbischen muss weitergeführt, die zum Obersorbischen sollte intensiviert werden. Denn das langfristige Ziel kann nur eine möglichst umfassende Aufbereitung des gesamten sorbischen Schrifttums sein.

Literatur

BARTELS, Hauke 2008: Passivkonstruktionen und Purismus: Konkurrierende Passivkonstruktionen in der niedersorbischen Schriftsprache. Ein Beispiel für Sprachwandel durch Purismus, in: KEMPGEN, Sebastian et al. (Hgg.), Deutsche Beiträge zum 14. Internationalen Slavistenkongress Ohrid 2008. München (= Die Welt der Slaven. Sammelbände; 32), S. 27–38.

BARTELS, Hauke 2010: Das (diachrone) Textkorpus der niedersorbischen Schriftsprache als Grundlage für Sprachdokumentation und Sprachwandelforschung, in: HANSEN, Björn; GRKOVIĆ-MAJOR, Jasmina (Hgg.), Diachronic Slavonic Syntax. Gradual Changes in Focus. München-Berlin-Wien, S. 7–18.

BARTELS, Hauke 2012: Maßnahmen zur Dokumentation des Niedersorbischen, in: Slavia Occidentalis 69, S. 7–22.

[99] Vgl. das Konzept von Texten als „Wissensrohstoff“ (HEYER/QUASTHOFF/WITTIG 2012).

BARTELS, Hauke 2013: Zur Konzeption eines historisch-dokumentierenden Wortschatz-Informationssystems des Niedersorbischen. Pläne zur Behebung eines drängenden Forschungsdesiderats, in: KEMPGEN, Sebastian; WINGENDER, Monika; FRANZ, Norbert; JAKIŠA, Miranda (Hgg.), Deutsche Beiträge zum 15. Internationalen Slavistenkongress, Minsk 2013. München-Berlin-Washington/D.C., S. 37–46.

BERRY, David M.; FAGERJORD, Anders 2017: Digital Humanities. Knowledge and Critique in a Digital Age. Cambridge-Malden MA.

BLESSING, André; KLICHE, Fritz; HEID, Ulrich; KANTNER, Cathleen; KUHN, Jonas 2015: Computerlinguistische Werkzeuge zur Erschließung und Exploration großer Textsammlungen aus der Perspektive fachspezifischer Theorien. DOI: 10.17175/sb001_013, in: Zeitschrift für digitale Geisteswissenschaften, Sonderband 1: Grenzen und Möglichkeiten der Digital Humanities, hg. von Constanze Baum und Thomas Stäcker. DOI: 10.17175/sb01.

BUBENHOFER, Noah 2009: Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. Berlin-New York.

BUBENHOFER, Noah; KUPIETZ, Marc 2018 (Hgg.): Visualisierung sprachlicher Daten. Heidelberg.

BUBENHOFER, Noah; ROTHENHÄUSLER, Klaus 2016: „Korporatheken“: Die digitale und verdatete Bibliothek, in: Zeitschrift für Bibliothekskultur 4/2, S. 60–71.

BUBENHOFER, Noah; SCHARLOTH, Joachim 2016: Kulturwissenschaftliche Orientierung in der Computer- und Korpuslinguistik, in: JÄGER et al. 2016, S. 924–933.

CARSTENSEN, Kai-Uwe; EBERT, Christian; EBERT, Cornelia; JEKAT, Susanne; KLABUNDE, Ralf; LANGER, Hagen (Hgg.) 2010: Computerlinguistik und Sprachtechnologie. Eine Einführung, 3., überarbeitete und erweiterte Auflage. Heidelberg.

DNW 2003–2020 – STAROSTA, Manfred; HANNUSCH, Erwin; BARTELS, Hauke (unter Mitarbeit von Fabian KAULFÜRST): Deutsch-niedersorbisches Wörterbuch. Internetversion (Vorabveröffentlichung, in ständiger Bearbeitung). Internet: https://www.niedersorbisch.de/dnw/.

EVERT, Stefan 2013: Tools for the Acquisition of Lexical Combinatorics, in: GOUWS et al. 2013, S. 1415–1432.

EVERT, Stefan; HARDIE, Andrew 2011: Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millennium, in: Proceedings of the Corpus Linguistics 2011 Conference. Birmingham. Internet: http://www.research.lancs.ac.uk/portal/services/downloadRegister/27713201/Paper_153.pdf [16.07.2020].

FASSKE, Helmut 1994: Der Weg des Sorbischen zur Schriftsprache, in: FODOR, István; HAGÈGE, Claude (Hgg.), Language Reform: History and Future, Vol. VI, S. 257–283.

FITSCHEN, Arne; GUPTA, Piklu 2008: Lemmatising and Morphological Tagging, in: LÜDELING/KYTÖ 2008, S. 552–564.

GABRIELETOS, Costas 2018: Keyness Analysis. Nature, Metrics and Techniques, in: TAYLOR/MARCHI 2018, S. 225–258.

GOUWS, Rufus H.; HEID, Ulrich; SCHWEICKARD, Wolfgang; WIEGAND, Herbert Ernst (Hgg.) 2013: Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume. Berlin-Boston (= Handbücher zur Sprach- und Kommunikationswissenschaft; 5.4).

GRAHAM, Shawn; MILLIGAN, Ian; WEINGART, Scott 2016: Exploring Big Historical Data. The Historian’s Macroscope. London.

Hagenbruch, André 2010: Flache Satzverarbeitung, in: Carstensen et al. 2010, Kap. 3.4, S. 264–279.

Heyer, Gerhard; Quasthoff, Uwe; Wittig, Thomas 2012: Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse. Herdecke-Witten.

Hirschmann, Hagen 2019: Korpuslinguistik. Eine Einführung. Berlin.

Hunston, Susan 2008: Collection Strategies and Design Decisions, in: Lüdeling/Kytö 2008, S. 154–168.

Jäger, Ludwig; Holly, Werner; Krapp, Peter; Weber, Samuel; Heekeren, Simone (Hgg.) 2016: Sprache – Kultur – Kommunikation. Ein internationales Handbuch zu Linguistik als Kulturwissenschaft. Berlin-Boston (= Handbücher zur Sprach- und Kommunikationswissenschaft; 43).

Jänicke, Stefan; Franzini, Greta; Cheema, Muhammad F.; Scheuermann, Gerik 2015: On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges, in: Borgo, Rita; Ganovelli, Fabio; Viola, Ivan (Hgg.), Eurographics Conference on Visualization (EuroVis). DOI: 10.2312/eurovisstar.20151113.

Jannidis, Fotis; Kohle, Hubertus; Rehbein, Malte (Hgg.) 2017: Digital Humanities. Eine Einführung. Stuttgart.

Kämper, Heidrun 2007: Linguistik als Kulturwissenschaft. Am Beispiel einer Geschichte des sprachlichen Umbruchs im 20. Jahrhundert, in: Kämper, Heidrun; Eichinger, Ludwig M. (Hgg.), Sprach-Perspektiven. Germanistische Linguistik und das Institut für Deutsche Sprache. Tübingen, S. 419–439.

Kämper, Heidrun 2015: ‚Kollektives Gedächtnis' als Gegenstand einer integrierten Kulturanalyse. Kulturlinguistische Überlegungen am Beispiel, in: Kämper, Heidrun; Warnke, Ingo H. (Hgg.), Diskurs – interdisziplinär. Zugänge, Gegenstände, Perspektiven. Berlin-Boston, S. 161–188.

Kaulfürst, Fabian 2010: Eksemplar laży w Americe. Knigły z lět 1727, in: Rozhlad 61/11, S. 17–18.

Kaulfürst, Fabian 2014a: Praktyczny przewodnik po korpusie języka dolnołużyckiego, in: Hebal-Jezierska, Milena (Hg.), Praktyczny przewodnik po korpusach języków słowiańskich. Warszawa, S. 67–75.

Kaulfürst, Fabian 2014b: Praktyczny przewodnik po korpusie języka górnołużyckiego, in: Hebal-Jezierska, Milena (Hg.), Praktyczny przewodnik po korpusach języków słowiańskich. Warszawa, S. 76–81.

Klingner, Jens; Lühr Merve (Hgg.) 2019: Forschungsdesign 4.0. Datengenerierung und Wissenstransfer in interdisziplinärer Perspektive. Dresden (= ISGV digital; 1). Internet: https://doi.org/10.25366/2019.04 [02.06.2020].

Kupietz, Marc; Schmidt, Thomas (Hgg.) 2018: Korpuslinguistik. Berlin-Boston.

Lasch, Alexander 2014: Zur Vereinbarkeit von diskurslinguistisch motivierter Sprachgeschichtsschreibung und maschineller Sprachanalyse am Beispiel des „Islamismus"-Diskurses, in: Vilmos Ágel; Gardt, Andreas (Hgg.), Paradigmen der aktuellen Sprachgeschichtsforschung. Berlin-Boston (= Jahrbuch für germanistische Sprachgeschichte; 5), S. 231–249.

Lüdeling, Anke; Kytö, Merja (Hgg.) 2008: Corpus Linguistics. An International Handbook, Vol. 1. Berlin-New York. (= Handbücher zur Sprach- und Kommunikationswissenschaft; 29.1).

Mair, Christian 2018: Erfolgsgeschichte Korpuslinguistik? Überlegungen zum Fortschritt in der Sprachwissenschaft, in: Kupietz/Schmidt 2018, S. 5–25.

MASTERPLAN ZEITUNGSDIGITALISIERUNG 2017 – Empfehlungen zur Digitalisierung historischer Zeitungen in Deutschland (Masterplan Zeitungsdigitalisierung). Ergebnisse des DFG-Projektes „Digitalisierung historischer Zeitungen", Pilotphase 2013–2015. Dresden, 29. Januar 2016/Berlin, 12. Juni 2017. Internet: https://www.zeitschriftendatenbank.de/zeitungsdigitalisierung/ [09.04.2020].

MENZEL, Thomas; POHONTSCH, Anja [im Druck]: Sorbisch, in: PLEWNIA, Albrecht; BEYER, Rahel (Hgg.), Handbuch der Sprachminderheiten in Deutschland. Tübingen.

REICHMANN, Oskar 2012: Historische Lexikographie. Ideen, Verwirklichungen, Reflexionen an Beispielen des Deutschen, Niederländischen und Englischen. Berlin-Boston.

RUNDELL, Michael; ATKINS, Beryl T. Sue 2013: Criteria for the design of corpora for monolingual lexicography, in: GOUWS et al. 2013, S. 1336–1343.

SCHARLOTH, Joachim 2018: Korpuslinguistik für sozial- und kulturanalytische Fragestellungen. Grounded Theory im datengeleiteten Paradigma, in: KUPIETZ/SCHMIDT 2018, S. 61–80.

SCHARLOTH, Joachim; EUGSTER, David; BUBENHOFER, Noah 2013: Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn, in: BUSSE, Dietrich; TEUBERT, Wolfgang (Hgg.), Linguistische Diskursanalyse: neue Perspektiven. Wiesbaden, S. 345–380.

SCHMIDT, Helmut 2008: Tokenizing and Part-of-Speech Tagging, in: LÜDELING/KYTÖ 2008, S. 527–551.

SCHREIBMAN, Susan; SIEMENS, Ray; UNSWORTH, John (Hgg.) 2016: A New Companion to Digital Humanities. Chichester.

SCHRÖTER, Juliane; TIENKEN, Susanne; ILG, Yvonne; SCHARLOTH, Joachim; BUBENHOFER, Noah (Hgg.) 2019: Linguistische Kulturanalyse. Berlin-Boston.

SINCLAIR, John 2004: Trust the Text. Language, Corpus and Discourse. London.

SKL 2014 – SCHÖN, Franz; SCHOLZE Dietrich (Hgg.), Sorbisches Kulturlexikon. Bautzen.

STEFANOWITSCH, Anatol 2020: Corpus Linguistics. A Guide to the Methodology. Berlin. DOI:10.5281/zenodo.3735822.

SZCZEPAŃSKI, Marcin 2012: Modelowanie danych do słownika elektronicznego na przykładzie artykułów słownika A. Muki, in: MOTORNYJ, V.; SCHOLZE, D. (Hgg.), Pytannja sorabistyky. Prašenja sorabistiki. XIII Mižnarodnyj sorabistyčnyj seminar 2010. L'viv-Budyšin, S. 26–61.

TARP, Sven 2012: Online Dictionaries: Today and Tomorrow, in: Lexicographica 28, S. 253–267.

TAYLOR, Charlotte; MARCHI, Anna (Hgg.) 2018: Corpus Approaches to Discourse. A Critical Review. London-New York.

WÖLKOWA, Sonja 2013: Hornjoserbski tekstowy korpus w nowej formje, in: Serbska šula 66/2, S. 44–47.

WÖLKOWA, Sonja 2014: Tekstowy korpus a dalše informaciske srědki wo hornjoserbskej rěči w interneće, in: Studia z Filologii Polskiej i Słowiańskiej 49, S. 59–71.