Hauke Bartels / Fabian Kaulfürst /
Marcin Szczepański / Sonja Wölke

Das Monitoring des sorbischen Schrifttums

Grundlagen und erster Jahresbericht eines neuen Forschungsvorhabens[1]

1. Einführung

Im Folgenden wird ein neues, seit 2018 vorbereitetes Forschungsprojekt des Sorbischen Instituts vorgestellt.[2] Das Vorhaben mit dem offiziellen Titel „Ständiges Monitoring des obersorbischen und niedersorbischen Schrifttums“ hat zum Ziel, einen wesentlichen Teil aller in einem Kalenderjahr erschienenen sorbischen Druckschriften in digitaler Form zu hochwertigen Korpustexten[3] aufzubereiten und diese anschließend hinsichtlich des dort verwendeten Wortschatzes zu untersuchen – und dies über viele Jahre, nach Möglichkeit dauerhaft. Der vorliegende Aufsatz liefert grundlegende Informationen zu diesem Projekt sowie einen ersten Bericht mit ausgewählten Ergebnissen der Analyse des Jahrgangs 2019.

Im Internet-Rechtschreibduden wird die Bedeutung von „Monitoring“ mit (Dauer-) Beobachtung (eines bestimmten Systems) angegeben.[4] Der Begriff leitet sich vom engl. Verb *monitor* ‚beobachten, kontrollieren‘ ab. Die optionale Bedeutungskomponente (Dauer) in der Bedeutungsbeschreibung des Duden findet sich im offiziellen Projekttitel als einleitendes Attribut (S t ä n d i g e s Monitoring […]). Damit kann im Hinblick auf Schrifttum als Objekt eines Monitorings nicht gemeint sein, dass ständig und umfassend beobachtet wird, sondern vielmehr, dass eine Beobachtung auf Dauer, über einen längeren Zeitraum angelegt sowie durch Regelmäßigkeit gekennzeichnet ist, sodass ein umfängliches Bild entsteht. Der Projekttitel benennt das Beobachtungsobjekt als Summe des niedersorbischen und obersorbischen Schrifttums, was freilich noch nach einer genaueren

---

[1] Die Autorinnen und Autoren des Artikels haben schwerpunktmäßig zu folgenden Abschnitten des Gesamttextes beigetragen: H. Bartels: Kap. 1, 2, 4.1, 4.4, 4.5, 5 sowie Gesamtredaktion; F. Kaulfürst: Kap. 4.2.1, 4.3.2 und 4.4; M. Szczepański: Kap. 3.1, 3.2.1, 3.2.2 sowie Datenvoranalyse Niedersorbisch; S. Wölke: 3.2.3, 4.2.2 und 4.3.1. Dieser erste Bericht enthält auch allgemeine Informationen zum Schrifttumsmonitoring und richtet sich an ein breiteres, nicht nur sorabistisches Publikum.

[2] Das Vorhaben wurde von Mai bis Dezember 2018 als Konzeptionsphase und in den Jahren 2019 und 2020 als Pilotphase durch eine Teilzeitstelle gefördert von der Stiftung für das Sorbische Volk, die jährlich auf der Grundlage der beschlossenen Haushalte des Deutschen Bundestages, des Landtages Brandenburg und des Sächsischen Landtages Zuwendungen aus Steuermitteln erhält. Das Projekt kann ab 2021 zunächst in einer dreijährigen Konsolidierungsphase fortgesetzt werden.

[3] Das im Monitoring zur Aufbereitung der Datengrundlage angewendete Verfahren ist ausführlich in Bartels 2020 beschrieben. Im vorliegenden Beitrag wird möglichst auf diesen Artikel verwiesen, um Wiederholungen zu vermeiden. Ansonsten s. Kap. 3.

[4] Internet: https://www.duden.de/rechtschreibung/Monitoring bzw. Rechtschreibduden 2020.

Definition verlangt: Es wird nicht „das Schrifttum“ in seiner abstrakten und vagen Ganzheit beobachtet, sondern das lexikalische Teilsystem des Nieder- und Obersorbischen (der Wortschatz, die Lexik), wie es sich in gedruckten – aber für die Analyse in digitaler Form vorliegenden – Texten eines bestimmten Zeitraums darstellt (dazu mehr in Kap. 3.1).

Eine ausführlichere und stärker auf Dynamik ausgerichtete Definition von Monitoring lautet: „Monitoring ist die Überwachung von Vorgängen. Es ist ein Überbegriff für alle Arten von systematischen Erfassungen (Protokollierungen), Messungen oder Beobachtungen eines Vorgangs oder Prozesses mittels technischer Hilfsmittel oder anderer Beobachtungssysteme.“[5] Wie noch darzustellen ist, ergibt sich in unserem Fall der Vorgangscharakter (= Entwicklung des Wortschatzes) aus der Folge von Einzelbeobachtungen (= „Protokoll“ eines Jahrgangs des Schrifttums), die bei längerfristiger Umsetzung des Vorhabens einen Vergleich zwischen den Einzelbeobachtungen und damit die Beschreibung einer Entwicklung über die Jahre ermöglicht. Die Systematik des Monitorings muss sich aus der angewandten Verfahrensweise ergeben (s. Kap. 3.2). Und dass auch beim Schrifttumsmonitoring „technische Hilfsmittel“ zum Einsatz kommen, wird im Folgenden deutlich werden.

Das hier vorgestellte Forschungsvorhaben lief 2019/20 zunächst in einer Pilotphase. Aufgabe war die praktische Erprobung, Umsetzung und Weiterentwicklung eines zuvor in groben Zügen entworfenen Verfahrens zur lexikalischen Analyse des ober- und niedersorbischen Schrifttums der Jahre 2019 und 2020. Es ist vorgesehen, jährliche Berichte zu Aspekten der lexikalischen Entwicklung der beiden sorbischen[6] Schriftsprachen zu veröffentlichen. Das Schrifttumsmonitoring zeitigt aber auch darüber hinaus vielfachen Nutzen.

Nach dem folgenden Kapitel 2 zur Genese und Begründung sowie zu den Zielen des Vorhabens wird in Kapitel 3 über die Datengrundlage und das Verfahren zu dessen Aufbereitung und Analyse informiert. Das umfangreiche Kapitel 4 bietet erstmals einen Auswertungsbericht, im vorliegenden Text zum Jahrgang 2019. Den Abschluss bildet eine erste Zwischenbilanz der Pilotphase.

2. Vorgeschichte, Begründung und Ziele des Vorhabens

2.1 Zur Genese des Projekts

Die Einrichtung eines Schrifttumsmonitorings gerade jetzt anzugehen, ergab sich aus zwei Entwicklungen. Zum einen war 2017/18, als die Überlegungen dazu konkreter wurden, bereits absehbar, dass die notwendigen technischen und methodischen Grundlagen für ein solches korpustextbasiertes Unterfangen in der Sorabistik in absehbarer Zukunft geschaffen sein könnten.[7] Wie bereits in der Einleitung erwähnt, ist der Einsatz technischer Hilfsmittel in Monitoringverfahren üblich und für eine umfassende und systematische „Beobachtung“ selbst des vergleichsweise wenig umfänglichen sorbischen Schrifttums unumgänglich.

---

[5] Internet: https://de.wikipedia.org/wiki/Monitoring [08.01.2021; Hervorhebungen H. B.].

[6] Der Begriff „sorbisch“ wird im gesamten Text als Oberbegriff für Nieder- und Obersorbisch verwendet.

[7] Zur Darstellung des aktuellen Standes s. Bartels 2020.

Zum anderen ergab sich 2017 zwischen der sprachwissenschaftlichen Abteilung des Sorbischen Instituts (SI) und dem Domowina-Verlag (LND) eine fruchtbare Diskussion über die Zukunft des Arbeitsfeldes „Orthografie(wörterbuch)". Anlass hierfür waren Vorbereitungen für die siebte Auflage des obersorbischen „Prawopisny słownik"[8] und ein im SI erarbeitetes Gutachten dazu.[9] Da auch das Sorbische Institut auf diesem Feld tätig ist – vor allem durch die laufenden lexikografischen Projekte sowie die Entwicklung von Datenbeständen und Anwendungen zur automatischen Rechtschreibkontrolle – stellte sich zudem die Frage nach dem zukünftigen methodischen Vorgehen, etwa mit Blick auf die Neuaufnahme oder Herausnahme[10] (vgl. Graf 2020) von Lemmata bei einem stets relativ „aktuellen" Rechtschreibwörterbuch. Ebenso wichtig ist eine anzustrebende Parallelität oder zumindest Kompatibilität von gedruckten und digitalen Ressourcen.

Für eine mögliche Neuordnung des Arbeitsfeldes wurde im August 2017 vom SI ein Konzept vorgelegt. Eine wesentliche Neuerung sollte demnach mit Blick auf eine zukünftige achte Auflage bzw. Neuausgabe des obersorbischen Rechtschreibwörterbuchs und damit verbundener Ressourcen die Schaffung einer „wohldefinierten, möglichst objektiven Datengrundlage" für die Herstellung und Aktualisierung orthografischer Hilfsmittel sein. Zu diesem Ziel trägt das hier vorgestellte Schrifttumsmonitoring, das im weiteren Verlauf der Beratungen vereinbart wurde, wesentlich bei.

Wichtige Voraussetzung war aber zunächst eine regelmäßige Lieferung der im Domowina-Verlag gedruckten Schriften[11] in digitaler Form zur weiteren Verarbeitung im SI. Ersteres wurde im Februar 2019 zwischen LND, SI und Stiftung für das Sorbische Volk vertraglich vereinbart und findet seit März 2019 regelmäßig statt. Das zur Weiterverarbeitung notwendige komplexe Verfahren, in dessen Verlauf das gesamte einbezogene Schrifttum zu hochwertigen computerlesbaren Korpustexten aufbereitet wird, wurde vom SI bis zur Einsatzreife entwickelt (s. Bartels 2020).

---

[8] 2014 erschien die „6., durchges." Auflage des unter diesem Titel veröffentlichten obersorbisch-deutschen Wörterbuchs von Pawoł Völkel, bearbeitet von Timo Meškank (PS 2014), das gleichzeitig den Grundstock der automatischen Rechtschreibkontrolle für Obersorbisch bildet und über die Webseite www.soblex.de auch online zugänglich ist. Für das Niedersorbische existiert kein Rechtschreibwörterbuch. De facto übernimmt diese Funktion das 1999 erschienene Niedersorbisch-deutsche Wörterbuch (Starosta 1999). Dieses ist seit 2012 auch online zugänglich (Rubrik „Niedersorbisch-deutsche Wörterbücher" auf dem Sprachportal www.niedersorbisch.de) und wurde dort 2019 an die aktuell geltende Rechtschreibung angepasst. Eine aktualisierte Druckausgabe ist in Vorbereitung. Darüber hinaus wurde 2019 für das Niedersorbische eine zentrale Suche für sorbische Wörter aus Starosta 1999, dem DNW (2003 ff.) sowie zusätzlichen Lexemen aus einer lexikalischen Datenbank, die als Grundlage für die Applikation zur automatischen Rechtschreibprüfung dient, bereitgestellt. Damit steht online erstmals ein Quasi-Rechtschreibwörterbuch für das Niedersorbische bereit, das jedoch noch nicht voll ausgebaut und in sich mikrostrukturell konsistent ist – so fehlen z. B. zum Teil die sonst üblichen Angaben zur deutschen Bedeutung. Die Seite ist zugänglich über https://www.niedersorbisch.de/ortografija/slownik.

[9] Anja Pohončowa: Posudk manuskripta za prawopisny słownik (7. nakład), oktober 2017.

[10] Zu den Wörtern, die in der 2020 erschienenen 28. Auflage des Rechtschreibdudens nicht mehr enthalten sind, gehört zum Beispiel das in der sorbischen Kultur prominente *Hochzeitsbitter*, os. *braška*, ns. *pódružba*, *póbratš*. In Bär (2020: 225) heißt es dazu erläuternd, sicher die sorbische Tradition nicht im Blick habend: „Denn wer heutzutage heiraten will, lädt die Gäste entweder selbst ein oder überlässt es einem professionellen *Hochzeitsplaner* oder *Wedding Planner*."

[11] Die im Domowina-Verlag erscheinenden Publikationen umfassen den weitaus größten Teil der Druckschriften in ober- und niedersorbischer Sprache.

2.2 „Neue Lexik“

Ein wichtiges Ziel des Schrifttumsmonitorings ist die Identifizierung „neuer Lexik“ im Sorbischen. Was ist damit gemeint? Rein verfahrenstechnisch – siehe die Darstellung des aktuellen Vorgehens bei der Analyse des Schrifttums in Kap. 3 sowie grundsätzlich in Bartels 2020 – bedeutet dies zunächst: nicht automatisch erkannte Wortformen in den Korpustexten. „Automatisch“, weil Grundlage des Verfahrens ein computergestützter Abgleich zwischen einem Wortformen-Lexikon (auch: lexikalische Datenbank; kurz: Lex-DB) und den in einem Korpustext vorkommenden Wortformen ist; „Wortformen“, weil in den Korpustexten konkrete Wortformen (auch: Textwörter, Tokens: z. B. ns. *słowoma*, *pśilubili*, *jogo*) auftreten, und nicht nur die Grundformen bzw. Lemmata, wie man sie etwa in Wörterbüchern sucht und findet (ns. *słowo*, *pśilubiś*, *wón*).

Bei „neuer Lexik“ handelt es sich daher zunächst um bisher im Wortformen-Lexikon nicht registrierte Formen. Dahinter können sich sehr verschiedene Phänomene verbergen. Zum einen geht es dabei um „fehlerhafte“ Formen in einem sehr weiten Sinne:[12]

1. Schreib- oder Druckfehler registrierter Wortformen (ns. *burksej* statt *burskej*, os. *bohotši* statt *bohatši*),
2. den Regeln nach mögliche, aber bisher in der Lex-DB nicht erfasste Schreibvarianten registrierter Wortformen (os. *hightech-firma* statt *hightechfirma*),
3. (durch Rechtschreibreform oder generell) veraltete oder dialektbasierte Schreibungen registrierter Wortformen, die als solche (aktuell normabweichende) nicht in der Lex-DB erfasst sind (ns. *snidanje* statt *snědanje*, os. *wurěkowanje* statt *wurjekowanje*),
4. nicht registrierte morphologische Varianten registrierter Wortformen (ns. *kralejstwo* statt *kralojstwo*, *španiski* statt *špański*, os. *Dortmundčan* statt *Dortmundźan*, *stoprocentowski* statt *stoprocentny*).

Die Identifizierung und „Kompensation“[13] dieser formalen Abweichungen ist zwar wichtig zur Herstellung qualitativ hochwertiger Korpustexte. Solchen Formen gilt aber nicht das Hauptinteresse des Verfahrens – obwohl aus bestimmten „Fehlern“ durchaus nützliche Schlussfolgerungen gezogen werden können (vgl. Kap. 4.4).

Eine zweite große Fallgruppe besteht aus Wortformen (und den entsprechenden Lexemen), die zwar (gut/längst) etabliert sind, aber in der Lex-DB, die sich im Wesentlichen auf die lexikalischen Bestände der in den letzten Jahrzehnten entstandenen Wörterbücher stützt, bislang nicht registriert sind. Einerseits ist dies dadurch bedingt, dass die für die Lex-DB ausgewerteten Wörterbücher bestimmte Lexeme nicht enthalten. Gründe hierfür gibt es verschiedene: Zufall (ein Wort wurde von den Lexikografen übersehen/vergessen), Purismus (z. B. ungeliebte Lehnwörter aus dem Deutschen) oder Verzichtbarkeit (z. B. mit Blick auf ein sorbisch-deutsches, in erster Linie auf Textverstehen ausgelegtes Wörterbuch wegen der leichten Erschließbarkeit der jeweiligen Bedeutung, so explizit in Starosta 1999[14]). Letzteres hat auch mit dem Zwang zur Sparsamkeit bzgl. der verfüg-

---

[12] Die angeführten Beispiele stammen aus dem analysierten Korpus des Schrifftums 2019.

[13] Damit ist gemeint, dass derartige Abweichungen von bereits registrierten normativen Wortformen zwar in den meisten Fällen im Korpustext erhalten bleiben, also nicht korrigiert werden, dass jedoch in den Annotationen die registrierte Normform ergänzt wird. So wird eine anschließende Lemmatisierung ermöglicht und damit ein Zugriff auf diese Textbelege auch über die Norm-Wortformen bzw. Lemmata.

[14] Im Vorwort heißt es, dass „aus Raumersparnis auf die Anführung von leicht erschließbarer Lexik wie Negationen, Partizipien, Verbalsubstantive weitgehend verzichtet wurde.“

baren Seitenzahl bei Druckausgaben zu tun. Dieses Motiv kann generell zur Nichtberücksichtigung von Lexik in gedruckten Wörterbüchern führen.[15] Außerdem spielt natürlich eine Rolle, welche lexikografischen Quellen für die beiden Lex-DB bereits ausgewertet werden konnten (vgl. hierzu den Beginn von Kap. 3).

Abgesehen davon, welche Wörterbücher für die Lex-DB ausgewertet wurden und was in ihnen enthalten ist, ist für das Schrifttumsmonitoring auch die Tatsache von Bedeutung, dass es bisher kein einziges sorbisches Wörterbuch gibt, das auf einer systematischen und umfassenden Auswertung des Schrifttums basiert. Bisherige, auch neuere Wörterbücher stützten sich auf eine partielle manuelle Exzerption von Literatur (so bei Starosta 1999 ebenso wie beim DOW 1989/91 oder dem DOW-NL 2006), ggf. punktuell ergänzt durch Korpusrecherchen (so beim DNW 2003 ff.). Seit Jahren wird daran gearbeitet, eine geeignete Datengrundlage für eine solche systematische Auswertung zu schaffen (Bartels 2020: Kap. 1 und 4.1–4.2). Ein erster Versuch, ein umfassendes korpusbasiertes historisches Wörterbuch zunächst für das Niedersorbische zu organisieren und über Drittmittel zu finanzieren (Bartels 2013), war letztendlich nicht erfolgreich. So gibt es leider bis heute keine umfassende, auf der vollständigen Auswertung des überlieferten Schrifttums basierende lexikografische Inventarisierung des sorbischen Wortschatzes. Das hier beschriebene Schrifttumsmonitoring stellt aber einen weiteren wichtigen Schritt in diese Richtung dar.

Eine besondere und quantitativ bedeutsame Gruppe der zu bearbeitenden Lexik bilden Eigennamen und von ihnen abgeleitete Formen. Siehe dazu Kap. 4.1. Besonders behandelt werden auch nicht-sorbischsprachige Teile in den analysierten Texten, siehe dazu Kap. 3.2.1 und 3.2.2.2.3.1 (Niedersorbisch) sowie Kap. 3.2.3.2 (Obersorbisch).

Von den oben genannten Gruppen „neuer“ Lexik werden zukünftig einige in die Lex-DB aufgenommen, ggf. in gesonderte „Sektionen“, um das künftige Verfahren effektiver zu gestalten. Dies kann auch einige Typen „fehlerhafter“ Formen betreffen, vor allem ist hier aber die zweite Fallgruppe der im Schrifttum etablierten, aber dennoch bisher nicht in der Lex-DB registrierten Lexeme relevant. Dies ist zugleich ein vorbereitender Schritt für die ausstehende lexikografische Inventarisierung des Sorbischen.

2.3 Neologismen

Neben den bisher angesprochenen Erscheinungen und abgesehen von den besonderen Fallgruppen der Eigennamen und der nicht-sorbischen Lexik geht es im Projekt vor allem um neue Wörter im engeren Sinne, d. h. um Neologismen. Hier muss jedoch zunächst zwischen verschieden weiten Auffassungen dieses Begriffs unterscheiden werden, für die u. a. folgende Fragen entscheidend sind:

1. Sollen auch neue Bedeutungen bereits bekannter lexikalischer Einheiten (Neubedeutungen)[16] einbezogen werden?

---

[15] Dies betrifft generell Lexeme aus hochproduktiven Wortbildungsprozessen, z. B. bestimmte Derivate wie Feminative und Deminutiva oder deadjektivische Abstrakta auf *-ość*. Dasselbe gilt auch für reihenbildende Komposita mit Zahlen wie *5-stronski*, *60-lětny* oder auch *dwanaćewosobowy*, *dwuetažowy* usw.

[16] Die Terminologie folgt hier Kinne 1998.

2. Sollen nur Wörter in engerem Sinne, d. h. Einwortlexeme, oder auch mehrteilige lexikalische Einheiten (Mehrwortlexeme) registriert werden?
3. Sollen auch nur einmalig oder sehr selten auftretende, möglicherweise als Gelegenheitsbildungen (Okkasionalismen, Ad-hoc-Bildungen) zu charakterisierende Wörter berücksichtigt werden?

Was die ersten beiden Fragen betrifft, so ist die Antwort mit Blick auf das Sorbische einfach. Wegen der verfügbaren Ressourcen (personell wie technisch) ist es derzeit schlicht nicht möglich, im hier beschriebenen Vorhaben neue Bedeutungen[17] und Mehrwortlexeme[18] systematisch und zuverlässig zu identifizieren. Daher müssen wir uns derzeit noch auf Neulexeme beschränken, die aus einem Textwort/Token[19] bestehen. Dabei ist auch eine endgültige Differenzierung nach Bildungsarten, nach denen mit Kinne (1998: 83) Neuschöpfungen, Neuprägungen/Neubildungen und Neuentlehnungen unterschieden werden können, nachgelagerten Analyseschritten im Zuge einer späteren lexikografischen Bearbeitung vorbehalten. Hinweise bzw. erste Einschätzungen dazu werden in der folgenden Darstellung gleichwohl gegeben.

Eine bekannte Definition für Neologismen, in der Bedeutungen eingeschlossen, Gelegenheitsbildungen jedoch ausgeschlossen werden, lautet wie folgt: „Ein Neologismus ist eine lexikalische Einheit bzw. eine Bedeutung, die in einem bestimmten Abschnitt der Sprachentwicklung in einer Kommunikationsgemeinschaft aufkommt, sich ausbreitet, als sprachliche Norm allgemein akzeptiert und in diesem Entwicklungsabschnitt von der Mehrheit der Sprachbenutzer über eine gewisse Zeit hin als neu empfunden wird.“ (Herberg/Kinne/Steffens 2004: XII)

Diese Definition setzt relativ strenge Maßstäbe an die Anerkennung eines in einem bestimmten Zeitabschnitt von der Sprachgemeinschaft als „neu“ empfundenen Lexems als Neologismus. Notwendig ist eine „Traditionskonstituierung“ (Kinne 1998: 77), die zugleich den Übergang eines Okkasionalismus, d. h. einer im alltäglichen Sprachgebrauch erfolgten spontanen lexikalischen Neuerung zum Neologismus markiert (ebd.: 78, 81).

Da ein Ausgangspunkt für die Organisation des Schrifttumsmonitorings Überlegungen zur Neukonzeption des obersorbischen Rechtschreibwörterbuchs waren, soll ergänzend das Vorgehen der Dudenredaktion angeführt werden, bevor die Möglichkeiten und Grenzen einer Neologismus-Registrierung im Sorbischen thematisiert werden. Auf der Internetseite des Dudenverlags[20] wird darüber informiert, wie „ein Wort in den Duden kommt“:

1. Ein Wort muss in einer „gewissen Häufung“ und in einer „bestimmten Streuung über die Texte hinweg“ auftreten, um ein „Neuaufnahmekandidat“ zu werden.

---

[17] Dies trifft auch auf das weiter unten erwähnte Projekt „Wortwarte“ zu.

[18] Hiermit sind lexikalische Einheiten (Lexeme) gemeint, die aus mehreren Textwörtern/Tokens bestehen, mithin lexembildende Wortgruppen wie z. B. dt. *kollektiver Freizeitpark* oder ns. *pisańske blido*, nicht jedoch z. B. Komposita wie ns. *nowopowědaŕ* oder Bindestrich-Schreibungen wie os. *ja-powědar*, die in der folgenden Analyse Berücksichtigung finden. Diese unterschiedliche Behandlung hängt damit zusammen, dass das derzeitige Verfahren sich noch auf einzelne Tokens beschränkt (vgl. dazu Bartels 2020: Kap. 4.4.2 und 4.4.7 sowie die folgende Fußnote).

[19] Im hier beschriebenen Vorhaben ist damit eine geschlossene Folge von Schriftzeichen im Text gemeint, die in der Mehrzahl der Fälle durch Leerzeichen abgegrenzt ist. Vgl. dazu Kap. 3.2.1.

[20] Internet: https://www.duden.de/ueber_duden/wie-kommt-ein-wort-in-den-duden [21.01.2021].

2. Die geforderte „Häufigkeit“ des Wortes muss „über einen längeren Zeitraum hinweg“ nachweisbar sein.
3. Das Wort sollte in verschiedenen Textsorten auftreten.

Die Messung der o. g. Kriterien wird auf Grundlage eines Textkorpus vorgenommen, das „mehr als 5 Milliarden Wortformen“ umfasst und 25 Jahre vorangehendes Schrifttum repräsentiert. Das – hier nur ausschnittweise wiedergegebene – Vorgehen der Dudenredaktion entspricht damit in etwa der Intention obiger Definition von Neologismen. Zugleich wird schnell deutlich, dass ein analoges Verfahren für das Sorbische nicht ohne Einschränkungen umsetzbar ist.

In der Lexikografie der Neologismen lassen sich grundsätzlich zwei Herangehensweisen unterscheiden. Es dominiert eine „retrospektive Neologismuslexikografie“ (Engelberg/Lemnitzer 2009: 58), wie sie logisch auch aus obiger Definition folgt – schließlich impliziert der Begriff Neologismus die historische Perspektive und damit eine Betrachtung von Sprachwandel. Denn ob ein neues Wort über längere Zeit auftritt und sich ausgebreitet hat (bzgl. Vorkommenshäufigkeit und Streuung), lässt sich erst im Nachhinein für einen definierten Zeitabschnitt beurteilen. Die Dudenredaktion betrachtet hier wohl in der Regel etwa ein Vierteljahrhundert, das Wörterbuch, aus dem obige Definition stammt, ein Jahrzehnt.

In Bezug auf das gerade erst begonnene Schrifttumsmonitoring ist klar, dass die Umsetzung eines solchen Verfahrens kurzfristig nicht möglich ist, langfristig aber eben gerade durch das hier beschriebene Vorhaben ermöglicht wird. Denn es erschafft – und das stellt ein weiteres wichtiges Ziel dar – durch die kontinuierliche Erarbeitung hochwertiger Korpustexte (vgl. Bartels 2020) ein in sich konsistentes, bzgl. der einbezogenen Quellen weitgehend lückenloses und damit valides Textkorpus beider sorbischer Schriftsprachen. Die letzten drei Auflagen des obersorbischen Rechtschreibwörterbuchs erschienen im Abstand von etwa neun Jahren (5. Auflage 2005, 6. Auflage 2014, die 7. Auflage wird voraussichtlich 2022 erscheinen), und wie oben dargestellt soll ein zeitgemäßes Verfahren zur Zusammenstellung der Lemmaliste ab der 8. Auflage, also Anfang/Mitte der 2030er-Jahre, verfügbar sein. Zu dem Zeitpunkt werden bei Fortsetzung des Monitorings bereits mehr als zehn Jahre Schrifttum analysiert sein, womit eine gute Datengrundlage zur Messung einer für sorbische Verhältnisse angemessenen Häufigkeit, Stetigkeit und Streuung neuer Wörter gegeben sein wird.

Warum „für sorbische Verhältnisse angemessen“? Aufgrund des vergleichsweise geringen Umfangs des obersorbischen und, noch stärker, des niedersorbischen Schrifttums muss ein praktikables Mindestmaß für die Kriterien Häufigkeit, Stetigkeit und Streuung des Auftretens eines Wortes im Korpus beim Schrifttumsmonitoring erst ermittelt werden. Neben der schon erwähnten Tatsache, dass wir bisher über kein aktuelles valides 10- oder 25-Jahres-Korpus verfügen, spielt in diesem Zusammenhang v. a. die verfügbare Korpusgröße eine wichtige Rolle. Das von der Dudenredaktion aktuell verwendete 5,6-Milliarden-Tokens-Korpus,[21] das lediglich 25 Jahre deutsche Schriftsprache repräsentiert, ist allein etwa 28-mal so umfangreich wie das gesamte sorbische Schrifttum vom 16. Jh. bis heute.[22] Aufgrund der für 2019 bereits vorliegenden Zahlen verfügen wir für diese erste

---

[21] Rechtschreibduden 2020: 9.

[22] Bei dieser Schätzung wird aufgrund der bisherigen Korpusgrößen für beide sorbische Schriftsprachen (vgl. Bartels 2020: Kap. 4.2) von einer hypothetischen Gesamtgröße des sorbischen Schrifttums von etwa 200 Millionen Tokens ausgegangen. Das dem erwähnten deutschen Neologismuswörterbuch zugrunde liegende Korpus umfasste 1,2 Milliarden Wortformen (Tokens) (Herberg/Kinne/Steffens 2004: XVI).

Jahrgangsanalyse über ein Korpus im Umfang von etwa 2,8 Millionen Tokens (vgl. Kap. 3.1). Dies ergibt für die Dekade 2019–28 bei annähernd gleichbleibender Textproduktion hochgerechnet eine Größe von ca. 30 Millionen,[23] für ein Viertel Jahrhundert ungefähr 75 Millionen Tokens. Beides liegt deutlich unterhalb der heutzutage als „Standard" betrachteten 100-Millionen-Grenze.[24] Da die Maßstäbe für Verfahren der Korpuslinguistik und Computerlexikografie am Beispiel „großer" Sprachen entwickelt wurden, ergibt sich für das sorbische Vorgehen die Notwendigkeit von Anpassungen und Pragmatismus.

Eine weitere für das Sorbische zu beachtende Besonderheit betrifft das Kriterium der Streuung, das sich zwar zunächst auf ein Textkorpus als Datengrundlage bezieht, aber wegen dessen Repräsentativitätsanspruchs (siehe z. B. Stefanowitsch 2020: 28 ff.) indirekt auch auf die Schriftsprache insgesamt. Wie bereits in Bartels 2010 für das Niedersorbische thematisiert, haben wir es im Sorbischen generell mit einer sehr geringen Diversität an Textsorten und Autor(inn)en zu tun. Dieser Befund wird weiter verschärft durch einen relativ großen Einfluss weniger Personen in ihrer Funktion als Verlagslektorinnen oder Korrektoren auf die konkrete sprachliche Gestaltung (vgl. zum Obersorbischen die Fußnote 110).

Das sorbische Schrifttumsmonitoring (wie generell die sorbische Lexikografie) muss sich die für das vorherrschende retrospektive Vorgehen notwendige Datenbasis erst erarbeiten. Es wird die Fähigkeit zur Umsetzung dieses Verfahrens bei Fortsetzung des begonnenen Monitorings aber schrittweise vergrößern und in etwa einem Jahrzehnt erreichen können.[25] Bis dahin, und vor allem solange einzelne oder nur wenige Jahrgänge analysiert werden können, entspricht das Vorgehen im sorbischen Schrifttumsmonitoring notwendigerweise eher einem zweiten Ansatz, den Engelberg/Lemnitzer (2009: 59) als „tagesaktuelle Neuwortlexikografie" bezeichnen. Ein solcher Ansatz wird z. B. im Projekt „Wortwarte"[26] verfolgt. Zwar

---

[23] Die jährliche Menge wird auf Grundlage der Zahlen aus 2019 mit drei Millionen Tokens angesetzt, da im ersten Jahr ein Teil der Textproduktion nicht geliefert wurde (vgl. Fußnote 29).

[24] So etwa das British National Corpus (nur 20. Jh., 100 Millionen Tokens; s. www.natcorp.ox.ac.uk, [08.01.2021]) oder das DWDS-Kernkorpus (20. Jh., mittlerweile ca. 120 Mio. Tokens; s. https://www.dwds.de/d/korpora/kern [08.01.2021]).

[25] Dieser Prozess ließe sich beschleunigen, indem das im ständigen Schriftumsmonitoring geplante Vorgehen durch eine nachträgliche Analyse des Schrifttums der letzten Jahrzehnte ergänzt würde, z. B. zunächst für die Jahre 2010–2018. Dies würde die Datengrundlage für weitergehende Analysen, Bewertungen und Beschreibungen schneller schaffen und damit zunächst das Monitoring fördern. Es würde darüber hinaus aber auch die Grundlage für eine empirisch stärker abgesicherte, d. h. korpuslinguistisch basierte zeitgenössische Lexikografie und Sprachbeschreibung bilden. Dies ist gerade für die Texte der vergangenen Jahrzehnte von großer Bedeutung, da sie – v. a. bedingt durch den politischen Umbruch 1989/90 – relativ starke Veränderungen in der sprachlichen Praxis widerspiegeln. Bisherige Versuche, ein solches schrittweises „retrogrades Monitoring" für das sorbische Schrifttum, idealerweise (zurück) bis 1945 zu finanzieren, waren jedoch nicht erfolgreich. Für das ältere (weitgehend gemeinfreie) niedersorbische Schrifttum erfolgt ein retrogrades Monitoring mittlerweile im Rahmen eines durch den Bund geförderten Drittmittelprojekts (s. Bartels 2020: Kap. 5.2.3).

[26] Internet: https://wortwarte.de [14.08.2020]. Ziel des Projekts ist eine zeitnahe Beobachtung von „Tendenzen der Entwicklung des Deutschen" im lexikalischen Bereich und damit gerade nicht erst solcher „neuer" lexikalischer Einheiten, die bereits über einen längeren Zeitraum verbreitet und anerkannt sind (s. Website sowie Engelberg/Lemnitzer 2009: 59). – Zum Zeitpunkt der Endredaktion dieses Artikels Anfang Februar 2021 fand ein Umzug des Projekts statt, das wohl einhergehend mit konzeptionellen Änderungen unter der Adresse wortwarte.org [09.02.2021] fortgesetzt wird.

ist statt einer Tagesaktualität[27] im sorbischen Schrifttumsmonitoring nur eine Jahresaktualität angestrebt, die grundlegenden Eigenschaften einer solchen nicht auf Mehr-Jahres-Zeiträume bezogenen Vorgehensweise sind aber vergleichbar.

So wie in den vergangenen zehn Jahren die Voraussetzungen für ein sorbisches Schrifttumsmonitoring geschaffen werden konnten, so kann in der kommenden Dekade die Grundlage für eine moderne, korpustextbasierte Lexikografie für die sorbischen Schriftsprachen gelegt werden. Dabei lassen sich manche Nachteile „kleiner Sprachen" durchaus auch in Vorteile wenden, da die relative „Kleinheit" des Schrifttums bestimmte Möglichkeiten erst eröffnet, so z. B. für den Aufbau eines „historischen Vollkorpus" oder für eine aufwendige und gründliche Aufbereitung von Korpustexten als Basis für lexikografische und viele andere Forschungen (vgl. Bartels 2020).

3. Datengrundlage und Vorgehensweise

Während der 2019–20 laufenden Pilotphase des Schrifttumsmonitorings wurde bereits versucht, ein für beide sorbische Sprachen möglichst einheitliches Vorgehen zu etablieren. Verschiedene Gründe führten aber dazu, dass bei diesem Vorhaben bis auf Weiteres noch keine vollständige Einheitlichkeit erreicht werden kann. Die Hauptursache hierfür liegt in den Unterschieden der lexikalischen Datenbanken (Lex-DB) für Nieder- bzw. Obersorbisch. Diese sind aus historischen Gründen auf unterschiedliche Weise und mit teilweise unterschiedlicher Zielstellung entstanden. Der Inhalt[28] wie auch die technisch-konzeptionelle Realisierung der beiden Lex-DB unterscheiden sich daher deutlich, was sich auf deren Nutzbarkeit wie auch auf personelle Anforderungen für deren Verwendung und Bearbeitung auswirkt. Die Differenzen im Vorgehen führen dazu, dass nicht immer für beide Sprachen direkt vergleichbare Ergebnisse vorgelegt werden können. Dieses Problem kann erst im Laufe der Jahre schrittweise gelöst werden. Die Erprobung unterschiedlicher Methoden bietet aber auch die Möglichkeit, durch den Vergleich ihrer Praktikabilität und Effektivität die Vorgehensweise zu optimieren.

So wie das im Folgenden beschriebene Verfahren in der Pilotphase erst entwickelt und erprobt werden musste – ein Prozess, der anschließend fortzusetzen ist – so ist auch das Format dieses Berichts als ein erster Versuch zur Präsentation ausgewählter Ergebnisse des angelaufenen Monitorings zu betrachten.

Zeitlich hoffen wir die Abläufe so organisieren zu können, dass Jahresberichte für ein Schrifttumsjahr jeweils in den ersten Monaten des übernächsten Jahres vorliegen werden – so wie dieser Bericht zum Jahr 2019 in der ersten Jahreshälfte 2021 erscheint. Diese

---

[27] Tatsächlich werden auf der Internetseite in geringen zeitlichen Abständen neue Wörter vorgestellt, so z. B. am 2. August 2020: *Biogemüsekiste*, *Digitalgejammer*, *Gesichtsaustausch*, *Maskensektor*, *Pop-up-Bikelane*, *Schwimmuhr* und *Zweierkonstruktion* [14.08.2020].

[28] Dies betrifft nicht nur einzelne Informationspositionen sowie die Art der Kodierung von Inhalten und Inhaltsrelationen, sondern auch den Umfang der zugrunde liegenden Daten. Ein wichtiges Manko der obersorbischen Lex-DB ist, dass sie noch nicht alle relevanten Daten des DOW 1989/90 umfasst. An der Behebung dieses Problems wird aktuell am SI gearbeitet. Als Zwischenergebnis wurde Ende 2020 eine Online-Fassung dieses umfangreichen Wörterbuchs über die Seite www.obersorbisch.de/dow bereitgestellt. Zudem konnte ein Teilbestand der exklusiv in diesem Wörterbuch gebuchten Lexik einschließlich der Flexionsformen bereits Anfang 2021 in die Lex-DB integriert werden. – Zur niedersorbischen Lex-DB siehe Bartels 2020: Kap. 4.4, zur obersorbischen unten in Abschnitt 3.2.3.1.

zeitliche Abfolge ist zurzeit nur unter großen Anstrengungen zu halten; Erleichterungen ergeben sich aber hoffentlich aus der Anreicherung der Lex-DB und der daraus resultierenden Reduktion der Anzahl nicht registrierter Tokens.

3.1 Datengrundlage

Wie oben dargestellt, wird die Textgrundlage für das Schrifttumsmonitoring aus den vom Domowina-Verlag gelieferten Drucktexten in digitaler Form gebildet, für diesen Bericht aus denjenigen des Erscheinungsjahres 2019. Die beiden folgenden Tabellen geben einen groben Überblick über diese Textbasis, genauere Angaben finden sich in der Anlage.[29]

Titel	Druckseiten	Textwörter[30]
Serbske Nowiny: Njewotwisny wječornik za serbski lud	1 073	1 589 919
Rozhlad: Serbski kulturny časopis (hier ohne niedersorbische Texte)	428	160 990
Sonstige	2 277	476 471
Gesamt	**3 778**	**2 227 380**

Tab. 1: Analysierte Textmenge in obersorbischer Sprache 2019

Titel	Druckseiten	Textwörter
Nowy Casnik: Tyźenik za serbski lud	332	410 211
Rozhlad: Serbski kulturny časopis (hier nur Artikel in niedersorbischer Sprache)	101	27 068
Sonstige	604	133 460
Gesamt	**1 037**	**570 739**

Tab. 2: Analysierte Textmenge in niedersorbischer Sprache 2019

3.2 Analyseverfahren

Einige der im Folgenden beschriebenen Schritte werden für beide sorbische Sprachen in einheitlicher Weise durchgeführt, bei anderen weicht die Vorgehensweise aus den oben

---

[29] Im ersten Jahr des Schrifttumsmonitorings 2019 wurden dem Institut nicht alle Texte zur Verfügung gestellt: So fehlen die ersten 48 Ausgaben der „Serbske Nowiny“ (SN) und die ersten elf Ausgaben des „Nowy Casnik“ (NC). Werbetexte aus SN, NC, dem „Rozhlad“ und der „Serbska pratyja“ werden grundsätzlich nicht geliefert. Auch ein Teil der Bildunterschriften (ca. 40 %) aus der „Serbska protyka“ fehlte. – In der Anlage mit vollständiger Quellenliste sind die einzelnen Texte nummeriert (z. B. Os-12, Ns-5). Diese Nummern dienen auch als Verweis im Text.

[30] Die hier gezählte Einheit „Textwörter“ ist nicht identisch mit der später gezählten Einheit „Tokens“, da die an dieser Stelle vorgenommene Zählung der eigentlichen Tokenisierung vorgelagert ist. Hier sind zunächst alle durch Leerzeichen getrennten Folgen von Buchstaben gemeint, ohne jede Bearbeitung. Damit werden z. B. auch deutsche und andere fremdsprachige Bestandteile der sorbischen Texte gezählt.

genannten Gründen ab. Dies spiegelt sich notwendigerweise in der Darstellung des Verfahrens und teilweise auch der Analyseergebnisse wider.

3.2.1 Vereinheitlichung der Textkodierung, Tokenisierung und Lemmatisierung

Die zu analysierenden Texte wurden in digitaler Form, aber in unterschiedlichen Formaten zur Verfügung gestellt, weshalb zunächst eine formale Vereinheitlichung notwendig war. Vorgehen und Formate entsprechen dabei dem in Bartels 2020 (Kap. 4.3) beschriebenen Verfahren. Im Ergebnis enthalten die Korpustexte alle notwendigen Metadaten und eine textstrukturelle Annotation in einem anerkannten Standardformat (TEI-P5). Diese Strukturierung ermöglicht auch einen effektiven Umgang mit verschiedensprachigen Passagen (ebd.: Kap. 4.4.6). Zur Vorbereitung der weiteren Textanalyse sind darüber hinaus eine Tokenisierung sowie eine Lemmatisierung notwendig (ebd.: Kap. 4.4).

Der Prozess der Tokenisierung zur Abgrenzung einzelner Textwörter (sog. Tokens) erfolgt grundsätzlich automatisch und für beide sorbische Sprachen einheitlich (ebd.: Kap. 4.4.2). Für Ober- und Niedersorbisch wird derselbe Tokenisierer genutzt, der allerdings auf Grundlage sprachspezifischer lexikalischer Datenbasen und diverser ergänzender Listen (z. B. für Abkürzungen) operiert.

Einige Ergebnisse bzw. Lösungsvorschläge des Tokenisierers müssen manuell korrigiert werden, z. B. mit Blick auf eine erwünschte bzw. unerwünschte Token-Trennung bei Verwendung von Schrägstrich (/) oder Trennstrich (-) im Text; im Ergebnis (die Tokengrenzen werden durch | markiert): |*Sorben-/Wendenrat*|, |*Chóśebuz*|/|*Cottbus*|, |*Maćica*|/|*Maśica*|, |*k-dispoziciji-byśe*|/|*wukubłanje*|, |*foto-wustajeńca*|, |*e-mail*|, |*Witajźiśi*|, |*wucabnik*|-|*wuknik*|, |*Schleswig-Holsteinska*|, |*Barliń*|-|*Bramborska*|, |*wał.*|-|*stw.*|, |*serbsko-nimsko-dańsk*i|, |*Atlas V-raketa*|, |*pśe-wa-li-jo-my*|. Einige dieser Korrekturen fließen anschließend in die lexikalischen Datenbasen ein (z. B. in Listen für „Ausnahmen“), um in Zukunft die Ergebnisse der automatischen Textanalyse zu verbessern und somit den Aufwand für manuelle Nachbearbeitung zu reduzieren. Auch offensichtlich fehlerhafte Schreibungen oder Worttrennungen werden soweit wie möglich behoben.

Bei der Lemmatisierung wird versucht, jedem zuvor identifizierten Token eine lexikalische Grundform (Lemma) zuzuweisen (siehe ebd.: Kap. 4.4.3). Mit Blick auf das Hauptziel des Schrifttumsmonitorings, die Identifizierung „neuer Wörter“, ist es wichtig, dass das Zuordnen von Tokens nicht nach abstrakten Algorithmen erfolgt, die z. B. auf einem bestimmten Grad formaler Ähnlichkeit beruhen, sondern durch einen Abgleich mit den lexikalischen Datenbanken (Lex-DB). Nur für die bereits in der Lex-DB verzeichneten Wortformen wird ein Lemma zugewiesen. Alle anderen, d. h. die nicht automatisch lemmatisierten Tokens gelten zunächst als „unbekannt“ und sind Gegenstand der weiteren Bearbeitung.[31] Da bei diesem Verfahren nur in der Lex-DB verzeichnete Flexionsformen erkannt und einem Lemma zugeordnet werden, werden so auch „neue“

[31] Ein gewisses Problem stellen dabei sog. „false positives“ dar, d. h. erfolgreiche Zuordnungen Textwort–Lemma, die sich bei näherer Prüfung als falsch erweisen, z. B. wenn der Eigenname *(Jan) Buk* aufgrund von Homonymie dem in der Lex-DB verzeichneten Lexem (Lemma) *buk* ‚Buche‘ zugewiesen wird. Der im Text belegte Familienname sollte stattdessen eigentlich als neues Lexem des Typs „Eigenname“ klassifiziert werden. Dieses Phänomen, das zu den erklärten Beschränkungen des derzeitigen Verfahrens zählt (siehe Bartels 2020: Kap. 4.4.7), muss zurzeit noch in Kauf genommen werden.

Flexionsformen bereits bekannter Lexeme identifiziert, z. B. ns. *kilomejtari* als bisher nicht registrierte Nebenform (Gen. Pl.) der in der Lex-DB verzeichneten Flexionsform *kilomejtarjow* des Lexems *kilomejtaŕ*.

Die zuvor durchgeführte Annotation größerer fremdsprachiger Fragmente (Absätze, Artikel oder ganze Seiten), die im Niedersorbischen etwa 13 % aller Tokens enthalten – vor allem aus deutschsprachigen Seiten im „Nowy Casnik" –, erlaubt deren Ausschluss aus der weiteren Analyse.[32]

3.2.2 Spezifika Cottbus/Niedersorbisch

3.2.2.1 Unerkannte Formen

Die nicht erfolgreich automatisch lemmatisierten Wortformen sind Gegenstand weiterer Behandlung, die zunächst der Verbesserung der Datenbasis für die spätere linguistische Bewertung dient. Im Wesentlichen geht es dabei um den Ausschluss von Daten, um die Menge der zu beurteilenden Fälle sinnvoll einzugrenzen (Kap. 3.2.2.2). Aber die als relevant deklarierte Fallmenge kann auch wieder anwachsen, indem in dieser Phase der Aufbereitung der Datengrundlage diverse „Fehler" korrigiert werden, die ansonsten zum ungewollten Ausschluss von Daten führen würden (Kap. 3.2.2.3).

Die folgende Tabelle gibt einen Überblick über die Mengenverhältnisse im zu diesem Bearbeitungszeitpunkt vorliegenden niedersorbischen Material:

Titel	Tokens in niedersorbischen Texten			
	Interpunktionszeichen	Wortformen		
		gesamt	autom. lemmatisiert	nicht autom. lemmatisiert
Nowy Casnik	64 006	337 667	303 274	34 393
Rozhlad	4 834	26 807	23 607	3 200
Serbska pratyja 2020	7 848	46 269	41 896	4 373
andere	13 800	82 783	76 300	6 483
gesamt	**90 488**	**493 526**	**445 077**	**48 449**

Tab. 3: Ergebnis der Tokenisierung und Lemmatisierung[33] niedersorbischer Texte 2019

3.2.2.2 Identifizierung nicht relevanten Sprachmaterials

Mit Blick auf die folgenden Ausführungen muss zunächst darauf hingewiesen werden, dass das derzeit angewendete Verfahren – ungeachtet der für Nieder- und Obersorbisch

---

[32] Zur Spracherkennung und -kennzeichnung in sorbischen Korpustexten s. ebd.: Kap. 4.4.6.

[33] Mit Blick auf die später in Tab. 5 für das Obersorbische angeführten Lemmatisierungsquoten: Die beim Niedersorbischen erreichte Quote lässt sich aus obiger Tabelle nicht ermitteln, da die Zahlen noch das im Folgekapitel 3.2.2.2 beschriebene „nicht relevante Sprachmaterial" umfassen. Die 2019 für das Niedersorbische erzielte Lemmatisierungsquote beträgt 96,25 %.

teilweise noch unterschiedlichen Vorgehensweise – überhaupt den ersten Versuch darstellt, große Mengen sorbischer Texte weitgehend automatisch zu analysieren. Traditionelle Methoden der Exzerption lexikalischen Materials aus Texten stützen sich wesentlich auf eine u. a. fachwissenschaftlich gegründete Intuition. Dabei werden Texte meist mit spezifischem Erkenntnisinteresse und bei – im menschenmöglichen Rahmen – möglichst gleichbleibender Aufmerksamkeit durchgesehen und „interessante" Fälle (z. B. „neue Wörter") exzerpiert. Das auf solche Weise zu analysierende Material ist quantitativ sicher begrenzt.[34] Wichtig für das Verständnis der in diesem Kapitel sowie im entsprechenden Abschnitt zum Obersorbischen beschriebenen „nicht relevanten" Fälle ist folgender methodischer Unterschied: Beim manuellen Exzerpieren kann man sich auf eindeutig „interessante" Fälle beschränken, während bei einer automatischen Analyse jeder einzelne Fall in irgendeiner Weise und nach möglichst klaren und transparenten Kriterien behandelt werden muss. Daher ist auch eine Beschäftigung mit Phänomenen notwendig, die nicht im eigentlichen Erkenntnisinteresse liegen, und zwar auch mit dem Ziel, den Aufwand dafür in zukünftigen Analyserunden zu reduzieren. Dies betrifft folgende Kategorien der nicht automatisch lemmatisierten Lexik.

3.2.2.2.1 Ziffern, Symbole und andere Nicht-Wörter

Die größte Gruppe für das Schrifttumsmonitoring nicht relevanter Tokens (ca. 15 000) besteht aus Formen, die man zusammenfassend als „Nicht-Wörter" bezeichnen kann. Dabei handelt es sich vor allem um:

- Tokens, die nur oder überwiegend aus Ziffern bestehen: *2004*, *½*, *14.650*, *01.09.2019*, *5:15*, *978-3-7420-2572-2*, *1536°C*, *7000m²*, *E.070a*. Aus dieser Gruppe werden einige Fälle gesondert behandelt und dennoch weiterbearbeitet, z. B. *10%ojsku*, *100lětne*, *1920tych*, *Bündnis90*;
- römische Zahlschrift: *IV*, *XIII*., *MDCI*;
- Webadressen, Hashtags, Dateinamen und andere „Codes" der digitalen Welt: *www.domowina-verlag.de*, *gromaze@gmail.com*, *dolnoserbski.de*, *#AppellvonCottbus*, *zur_Staerkung_der_niedersorbischen_Sprache__23.02.2019_.pdf*;
- technische Symbole: *€*, *©*, *°C*;
- Signaturen: *ChP*, *ChPi*, *J*, *Jš*, *pm*, *Ś.-ka*, *SchiDD*, *AvO*, *popcon*;
- morphologische, phonetische und ähnliche Elemente linguistischer Beschreibungen: *Zuk „ó"*; *pśez-*; *měke sycawki [ś], [ź], [tś]*; *słowjańske paršonowe mě na Rad-: Radoch abo Radoš*; *sanskritski kórjeń bhaj*; *wzejomy njewjericku, wótwóstajijomy nje-, pśidajomy wje-*;
- diverse andere: *ó=o*, *C+M+B*, *CO_2*, *B*, *b*.

3.2.2.2.2 Initialen, Abkürzungen und besondere Bezeichnungen

Am Rande der Kategorie der Nicht-Wörter liegen Initialen (*A.*, *W.*, *Ch.*, *Chr.*, *Sch.*, *Schm.*). Obwohl sie nahezu immer vollständige Textwörter repräsentieren – es handelt sich ja im Prinzip um Abkürzungen –, ist dennoch eine Zuordnung zu den „gemeinten" Vollformen im Rahmen des Projekts nicht möglich. Daher wird diese sehr frequente (ca. 1600 Tokens), aber für das Schrifttumsmonitoring nicht relevante Fallgruppe aus dem weiteren Verfahren ausgeschlossen.

---

[34] Die herkömmliche Exzerption basiert – bedingt durch unterschiedliches Ausgangswissen sowie durch einen unterschiedlichen Grad an Reflexion – auf subjektiver Einschätzung des jeweiligen Exzerptors.

Wegen ihrer besonderen Stellung und ebenfalls geringen Bedeutung für die sprachliche Analyse wurden ebenfalls folgende Kategorien (zusammen ca. 2600 Tokens) zunächst ausgesondert:

- (Noch) nicht etablierte, häufig ad hoc gebildete Abkürzungen: *ew.*, *fał.*, *febr.*, *mj.*, *proc.*, *st.*, *w.*, *wóž.*, *z.d.*;
- Akronyme: *AfD*, *CIOFF*, *KSW*, *SLŹ*, *SRWŠ*;
- Bezeichnungen biblischer Bücher: *Cef*, *Daniel*, *Eks.*, *Pśisł*, *sir*, *Tim.*, *tit*;
- physikalische Einheiten: *km/h*, *km²*, *m³*, *MHz*, *R*;
- einige Bezeichnungen aus dem Kirchenjahr: *epifanias*, *kantate*, *palmarum*, *septuagesimae*, *trinitatis*.

3.2.2.2.3 Nicht-niedersorbische und nicht normgerechte Lexik

Analyseobjekt des Schrifttumsmonitorings sind aktuelle sorbische Drucktexte in digitaler Form. Daher wird zunächst davon ausgegangen, dass die Sprache dieser Texte ein zeitgenössisches Nieder- oder Obersorbisch repräsentiert, das den geltenden orthografischen, grammatischen und lexikalischen Normen entspricht. Auf diesen Normen[35] basiert auch der Kern der lexikalischen Datenbanken, die als Referenzsystem für die automatische Textanalyse dienen. Gleichzeitig ist klar, dass in den gelieferten Texten keine vollständige sprachliche Synchronität und Homogenität erwartet werden kann. Das im Monitoring angewandte Verfahren erfordert daher eine Begrenzung des eingehender zu analysierenden Materials auch in dieser Hinsicht. Im Folgenden werden weitere Kategorien von Wortformen aufgeführt, die aus diesem Grund von der weiteren Analyse ausgeschlossen sind. (Der in diesem Verfahrensschritt erfolgende Ausschluss nicht normgerechter Phänomene kann sich aber nur auf klare Fälle von Abweichung bzw. entsprechender Nicht-Relevanz beziehen und daher nicht vollständig sein. Daher finden sich später auch Ausführungen zu nicht-standardsprachlicher Lexik.)

3.2.2.2.3.1 Nicht-niedersorbische Lexik

Anders als im obersorbischen Material wurde beim Niedersorbischen versuchsweise eine Bestimmung der Herkunftssprachen der Tokens vorgenommen, sodass hier ein Eindruck von den quantitativen Verhältnissen gegeben werden kann. Insgesamt wurden 8674 Wortformen nicht-niedersorbischer Herkunft identifiziert, die nicht eindeutig und vollständig in das niedersorbische Sprachsystem integriert sind. Das folgende Diagramm gibt einen Überblick über die Herkunft der aus der weiteren Behandlung ausgeschlossenen Wortformen.[36]

---

[35] Entsprechend den existierenden normsetzenden Wörterbüchern, Rechtschreibregeln und Beschlüssen der beiden Sprachkommissionen. Druckwerke sind dabei gegenüber den ständig aktualisierten Lex-DB bereits teilweise veraltet.

[36] An dieser Stelle der Analyse geht es schon nicht mehr um die zuvor erwähnten 13 % aller Tokens, die überwiegend von den deutschsprachigen Seiten im NC stammen, sondern um fremdsprachige Elemente in ansonsten eindeutig niedersorbischen Texten bzw. Textpassagen.

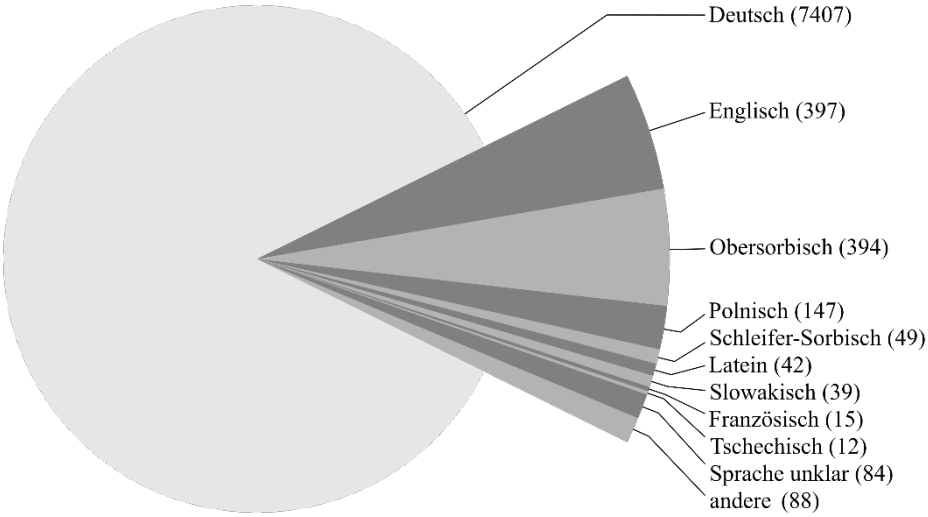

Diagramm 1: Nicht-niedersorbische Wörter im niedersorbischen Schrifttum 2019[37]

Nicht-niedersorbische Wortformen treten am häufigsten in folgenden Kontexten auf (in den Beispielen Formatierung wie im Original):

- In Zitaten: *„Měr buź z tobu – Friede sei mit dir“*; *Študańce w Chóśebuzu ›założichu sebi pod jeho wjednistwom prěnje delnjoserbske studentske towarstwo …‹*; *Aborigin jo wótegronił: „Kangaroo!“ (Ja śi njerozmějom!)*;
- in Titeln: *Engelske wudaśe ma titel „Shores of hope“*; *Za Kralowy »*Serbsko-němski słownik hornjołužiskeje rěče*« jo Šwjela 1933 pósudk pisał*; *Helmold w swójej „Chronica Slavorum“*;
- in Adressen: *Sukelnska 27, 02625 Budyšin, Tuchmacherstraße 27, 02625 Bautzen*; *na Spielbergowej droze 3 (Spielbergstraße)*;
- in Eigennamen: *Uniwersytet im. Adama Mickiewicza w Poznaniu*; *z Hermannstadta (Sibiu/Nagyszeben/Hermestatt)*; *Model „ta choćebuzska“ jo dostał prědne myto*;
- in metasprachlichen Verwendungen mit besonderer Kennzeichnung: *Pśeto »okno« – pšawje źe dejał rusojski groniś »akno« – jo teke na serbski »(w)okno«*, *Górnoserby gronje nejstaršej žonje w domje »maći«*, *na starem instrumenśe z mjenim „Nyckelharpa“*;
- teilweise auch einfach eingestreut: *w hinduizmje ważne słowo bhaktih*.

[37] Wortformen aus dem sog. Schleifer Sorbisch, d. h. aus dem dortigen Übergangsdialekt zwischen Nieder- und Obersorbisch, werden hier nicht einer der beiden Schriftsprachen zugeordnet und können daher auch nicht unter „Obersorbisch“ oder unter den Dialektphänomenen beider Schriftsprachen behandelt werden. Daher erscheinen sie hier als gesondere Fallgruppe. Die Gruppe der „anderen“ fasst Sprachen zusammen, aus denen maximal 10 Wortformen belegt sind: Dänisch, Hebräisch, Hindi, indigene Sprachen Nordamerikas, Irisch, Italienisch, Kroatisch, Maltesisch, Manx, Niederländisch, Norwegisch, Plattdeutsch, Rumänisch, Russisch, Schwedisch, unbestimmte slawische Sprache(n), Slowenisch, Spanisch, Stellingwerfs, Ungarisch, Westfriesisch.

In relativ vielen Fällen (1253) ist nicht leicht zu beantworten, in welchem Maße die in den Texten gebrauchten Wortformen ins niedersorbische Sprachsystem integriert sind. Häufig handelt es sich um Einzelbelege. Die folgenden Beispiele stehen für diese Gruppe, die in Zukunft noch Gegenstand genauerer Untersuchungen zu sprachlicher Interferenz und Graden der Integration sein könnten: *wó graśu Energie Cottbus pśeśiwo Bayern München*; *wědomnostnik wót Uni Jena*; *w Časopisu Maćicy Serbskeje*; *z direktorom tegdejšego Instituta za serbski ludospyt*; *pśiźćo do Gladhousa*; *spiw Berlinskeje dróhi*; *w pśewódniku pó Beskidzie Żywieckim*. Auch Belege ohne Eigennamen gehören in diese Gruppe: *zwězujotej minimal music a elektronisku pop muziku*; *póbitowanje ako all you can eat*; *wóni how njejsu sorbisch, ale wendisch*; *Serbski byś ako Cultural heritage w regionalnem sebjerozměśu Łužyce*; *Jo šło mjazy drugim wó tak pomjenjony permanent make up, to groni wó trajne šminkowanje*.

3.2.2.2.3.2 Alte und dialektale Lexik

Aus der zuvor nicht erfolgreich lemmatisierten Lexik wurden im Laufe der weiteren Analyse auch solche niedersorbischen Wortformen ausgeschlossen, die der geltenden Sprachnorm eindeutig nicht entsprechen, jedoch offensichtlich bewusst mit einer bestimmten Funktion im Text verwendet wurden. Dies betrifft Textwörter

- aus Zitaten alter Texte wie im folgenden Beleg: *[...] Wot Jordana, šulara Popojskego, roź. 1841 w Gornej Łužicy, (ma) Maśica swojo žyweńe jomu se źěkowaś*;
- aus alten Originaltiteln: *„Stronine meńa Chóśobuskego hobceŕeńa“*, *ako Bramborski Serski Casnik założony w lěše 1848* und
- explizit auf verschiedene Weise markierte Anführungen diverser dialektaler oder historischer Sprachformen: *Prossen (1412 Prossentin; wót paršonowego mjenja Prosęta, we kótarymž štycy słowo „pšosyś“)*; *pśeměnjenje e > a njejo se we wjelich słowach stało: wjeseliś se, wjecor, mjezy*; *źiwne zestajane słowa (na pśikład Schych Guy Zowsele; = Wšych gójcow zele; abo proday Zausele ‚brodajcow zele?‘).*

3.2.2.2.3.3 Stilisierungen und bewusste Normabweichungen

Eine besondere Gruppe nicht registrierter Lexik stammt aus längeren Texten, die zwar in zeitlicher Nähe zum Monitoringjahr 2019 geschrieben wurden, aber teilweise absichtlich von der geltenden Norm abweichen. Ein Beispiel: *Móžośo se hobźěliś na procesu hujasnjenja pšašanjow hokoło Bóžego blida EKBO a se k tomu až do kóńca maja 2019 hobźěliś na online-napšašowanju*. Solche Fälle werden ebenfalls ausgesondert.

3.2.2.3 Behebung von „Fehlern“ und Formatierungsdifferenzen

Zum Zweck einer möglichst zuverlässigen automatischen Analyse ist es notwendig, einige Tokens – auf Annotationsebene; vgl. Fußnote 13 – formal anzupassen. Dies betrifft uneinheitliche Kodierungen oder im Verfahren störende Formatierungsunterschiede, aber auch Druck- oder Schreibfehler, die leider in den Texten relativ häufig auftreten. Zum einen geht es um Großbuchstaben in Tokens, die am Satzanfang oder im ersten Wort eines Titels stehen, oder auch in Adjektiven als Teil von Eigennamen (*Nimsko-Pólske myto za žurnalisty → nimsko-pólske*) oder um Schreibungen in Versalien (*EUROPEAŹe → Europeaźe*, *WITAJ-KITA → WITAJ-kita*). Zum anderen ist es nötig, die Kodierung einiger Zeichen zu vereinheitlichen, um eine zuverlässige computergestützte Analyse zu

gewährleisten, z. B. den Apostroph[38] (*njeb´du → njeb'du*, *njeb`źoš → njeb'źoš*, *njam’žo → njam'žo*) oder Bindestriche (*serbsko–nimsku → serbsko-nimsku*). Ebenfalls werden (überwiegend schon bei der Tokenisierung) verschiedene „Interpunktionszeichen“ entfernt (*naźej[a] → naźeja*, *st wór jon a → stwórjona*, *pśewa-li-jo-my → pśewalijomy*), implizite Komposita aufgelöst (*Dolno- a Górnoserbow → Dolnoserbow a Górnoserbow*, *tu- a wukraja → tukraja a wukraja*) und expressive Formen verkürzt (*wuuu-śěgnjom → wuśěgnjom*, *ZMĚĚĚ-ROM → změrom*). Auch einfache Druck- oder Setzfehler werden korrigiert (*apryl,a → apryla*, *Bu dyšynje → Budyšynje*, *simultanjedo → simultanje do*).

Die Unterscheidung von mehr oder weniger zufälligen (druck)technischen und Schreibfehlern von systematischen oder gar bewusst herbeigeführten Abweichungen ist nicht immer einfach. Im bearbeiteten niedersorbischen Material erforderten ca. 1200 Tokens eine Korrektur. Die häufigsten Typen sind dabei mit über 200 Tokens falsche Buchstaben (*Sktualny → aktualny*, *pźez → pśez*, *zeleznizu → zeleznicu*), fehlende Buchstaben (*jazy → mjazy*, *miłosćim → miłosćiwym*, *pśdnosk → pśednosk*), zusätzliche Buchstaben (*Dolnoserbsam → Dolnoserbam*, *mjeńšynowyach → mjeńšynowych*, *tektsty → teksty*) sowie verschobene Buchstaben (*burksej → burskej*, *letkor → lektor*, *sersbka → serbska*).

Auf ähnliche Weise traten auch Fehler bei Buchstaben mit Diakritika auf (über 100 Tokens): *wśuźi → wšuźi*, *planowas → planowaś*, *droźe → droze*, *ceściła → cesćiła*. Eine genauere Analyse solcher „trivialer“ Fälle kann durchaus Hinweise auf Probleme der orthografischen Kodifizierung liefern. Einzelne Fallgruppen deuten z. B. hypothetisch auf Tendenzen in der Sprachentwicklung hin, so etwa auf Depalatalisierungstendenzen in gewissen phonetischen Kontexten.[39]

Einer genaueren Betrachtung bedürfen auch relativ häufige Schreibfehler folgender Typen (Beispiele):

- *r:ŕ* (ca. 140 Tokens):[40] *pjakarni → pjakaŕni*, *pśedsedarka → pśedsedaŕka*, *spiwarskich → spiwaŕskich*, *znajaŕjow → znajarjow*, *kantoŕka → kantorka*, *Brambořska → Bramborska*;
- *o:ó* (ca. 70):[41] *gnój → gnoj*, *pobył → póbył*, *škornje → škórnje*, *skrótka → skrotka*, *dóstawa → dostawa*, *wo → wó*;
- *e:ě:je* (30): *aktěry → aktery*, *kśe → kśě*, *nanejměnjej → nanejmjenjej*, *nějlěpjej → nejlěpjej*, *šěsć → šesć*, *śpje → śpě*, *trěbne → trjebne*[42];
- *l:ł* (30): *ałe → ale*, *służył → służył*, *złubił → zlubił*, *leśel → leśeł*, *młokom → mlokom*, *šło → šło*, *wuzjawil → wuzjawił*.

Weitere Fälle für eine spätere Analyse sind auch *wóśćy → wóśce*, *na žagly → na žagli*, *źěły → źěle*, *wótrostł → wótrosł*, *lažćej → lažčej*.

---

[38] Der Apostroph wurde mit dem Unicode-Zeichen „Apostrophe“ vereinheitlicht, um die Unterscheidung vom einfachen Anführungszeichen („Right Single Quotation Mark“) zu ermöglichen.

[39] Vgl. Kaulfürst (2019), wo in gewissen Fällen die Palatalität als fakultativ ausgewiesen ist: <[…] crjej> [[…] tsr(ʲ)e …] (S. 16), <[…] trějałko> [[…] ˈtr(ʲ)e(j)awkɔ] (S. 18).

[40] In den Beobachtungszeitraum fällt die Anfang 2019 erfolgte Verkündung der 2018 von der niedersorbischen Sprachkommission beschlossenen Ausweitung der Schreibung von <ŕ> im Wortinneren.

[41] In einigen Fällen, so bei *gnój*, *skrótka* und *dóstawa*, könnte es sich auch um Obersorabismen handeln (vgl. Kap. 4.3.2.6).

[42] Auch hier spielt eine Änderung der Kodifizierung aus dem Jahr 2017 eine Rolle, in der die Schreibung zu *trje-* vereinheitlicht wurde.

3.2.2.4 Datengrundlage für die weitere lexikalische Analyse

Als Ergebnis der bisherigen Schritte liegt annähernd die Menge Tokens vor, die in der Lex-DB nicht registrierte Wortformen repräsentiert. Es sei daran erinnert, dass es sich dabei nicht nur um Neologismen handelt, sondern um nicht registrierte und nur in diesem Sinne „neue“ Lexik ganz unterschiedlicher Art (siehe Kap. 2.2). Die folgende Tabelle zeigt die entsprechende Datenmenge für das Niedersorbische, schon nach bestimmten Kategorien geordnet:

Wortart	Tokens		Lexeme	
Verben	509	3 %	251	4 %
Partizipien	166	1 %	90	1 %
Substantive (Appellative)	2 192	12 %	1 074	17 %
Substantive (Eigennamen)	13 574	73 %	4 007	61 %
Adjektive	1 681	9 %	916	14 %
Weitere (Adverbien, Partikel, Kopulae, Pronomina)	367	2 %	166	3 %
Gesamt	**18 489**	**100 %**	**6 504**	**100 %**

Tab. 4: Anzahl bislang nicht registrierter Tokens bzw. Lexeme im Niedersorbischen

Das als für die weitere Analyse relevant eingestufte Sprachmaterial wurde grob geordnet und klassifiziert. Wie in der Tabelle teilweise schon erkennbar, wurden die Tokens bekannten oder neuen (bisher in der Lex-DB nicht enthaltenen) Lexemen zugeordnet und nach Wortarten und grammatischen Kategorien (Genus, Aspekt usw.) klassifiziert. Im weiteren Verlauf der Analyse wurden Normabweichungen typisiert (z. B als orthografisch, phonetisch oder morphologisch; als mögliche Schreibfehler, nicht registrierte Flexionsformen, Abweichungen in Affixen oder Stamm, bei der Wortbildung). Wo möglich wurde bei Abweichungen die erwartbare oder formal ähnliche Normvariante angeführt. Im Fall größerer Gruppen wurden weitere Kriterien beachtet, z. B. semantische Klassen bei Eigennamen. Auf diese Weise entstand eine relativ große Anzahl (ca. 160) Gruppen, die meist eine überschaubare Zahl von Wortformen umfassen. Jedem Token wurde zur Erleichterung der anschließenden Begutachtung automatisch der Verwendungskontext und die Quellenangabe hinzugefügt. Im Folgenden einige Beispiele für die Ergebnisse dieser Datenvoranalyse (die dortige Zählung hat nichts zu tun mit den hiesigen Kapitelnummern):

- 1-2-4: Verb » neues Lexem » ohne ein verwandtes Verb (16 Lexeme/47 Tokens): ⁻**twiterowaś** ip[43] — NC 12/5 *njewěźe, co se rownje **twiterujo** abo na instagramje wózjawjujo!*
- 2-2-2-1-3: Partizip » Partizip Präsens » Adverb » von einem erkennbaren Verb » von einem p-Verb (sic!) abgeleitet (8 Lexeme/9 Tokens): ⁻**chyśecy**, v. ⁺chyśiś p, erw. ⁺chytajucy, v. ⁺chytaś ip — NC 40/4 *wótrył, wusoko do lufta ju **chyśecy*** [erw. ⁺chytajucy]. *Publikum jo zajuskał*.[44]

[43] Die korrekte Schreibung wäre *twitterowaś*.

[44] „Verstöße“ gegen die aspektgesteuerte Bildung von Prozessualpartizipien sind auch für das Obersorbische häufig belegt, vgl. Kap. 4.2.2.2 (dort auch zum bestehenden Forschungsbedarf).

- 3-1-2-2: Substantiv » Appellativ » ohne auffällige Abweichung » ohne Bindestrich (530 Lexeme/1032 Tokens): ⁻**enkelsyn** m.anim, vgl. ⁺syn m — NC 15/1 *zgromadnje ze swójim* ***enkelsynom***. • NC 25/5 *žednu štrofu. Wóna jo z* ***enkelsynom*** *powědała a na jeje radu* • NC 49/2 *teke Dietrich Šwjela,* ***enkelsyn*** *fararja Bogumiła Šwjele* • NC 52/6 *se wjaseliła ak jeje* ***enkelsyn*** *jo pšišeł na tu serbsku*
- 3-2-3-1: Substantiv » Eigenname » anderer Name » Ethnie (57 Lexeme/92 Tokens): ⁻**Dolnonimc** m.anim — NC 27/2 *Flügge wót Towaristwa* ***Dolnonimcow*** *w Bramborskej jo rozpšawiła* • NC 27/2 *mógało rěcnu towarišnosć* ***Dolnonimcow*** *pšez to pódpěrowaś* • NC 35/8 *Lěwica: Strategiju* ***Dolnonimcam*** • NC 35/8 *až pódpěra pominanje* ***Dolnonimcow*** *w Bramborskej. Wóni* • NC 35/8 *part dalej wjasć a teke* ***Dolnonimcam*** *hyšći wěcej pomagaś.* • NC 35/8 *wěcej pomagaś. Měnje, až* ***Dolnonimce*** *w Bramborskej trjebaju* • NC 35/8 *zgromadnje ze zastupnikami* ***Dolnonimcow*** *na drogu spóraś.* • NC 35/8 *Flügge wót towaristwa* ***Dolnonimcow*** *w Bramborskej jo pšedstajiła*
- 4-3-7: Adjektiv » neues Lexem » andere Adjektive (240 Lexeme/374 Tokens): ⁻**kajakański** — NC 31/5 *małke wustawadliśćo* ***kajakańskeje*** *pšepóžycańe a parkowaniśćo*

3.2.3 Spezifika Bautzen/Obersorbisch

Wie bereits erwähnt, werden in der Pilotphase des Schrifttumsmonitorings (2019/20) verschiedene Vorgehensweisen erprobt. Unter anderem deshalb wurde zur Analyse der obersorbischen Korpustexte und anders als im Niedersorbischen die im SI entwickelte Software Corproc[45] eingesetzt. Dies und andere Faktoren bedingen diverse Unterschiede im Datenanalyseverfahren, die im Folgenden für das Obersorbische dargestellt werden. Im weiteren Verlauf des Projekts werden diese mit Blick auf das zukünftige Vorgehen bewertet.

3.2.3.1 Einsatz von Corproc für das Monitoring

Auch die obersorbischen Dateien wurden zunächst für das Monitoring aufbereitet und in ein einheitliches Format gemäß TEI-P5 gebracht. Die weitere Bearbeitung erfolgte jedoch direkt mit Hilfe von Corproc. Auch hier wurde zunächst automatisch tokenisiert. Notwendige Korrekturen fehlerhafter Tokengrenzen, wie in Kap. 3.2.1 beschrieben, können hier erst später, nach der automatischen Lemmatisierung in Corproc, erfolgen.[46]

Die dem Abgleich zugrunde liegende obersorbische lexikalische Datenbank basiert auf dem Lexikbestand und dem morphologischen Generator des zweisprachigen digitalen Wörterbuchs unter www.soblex.de.[47] Es folgte, wie in Corproc vorgesehen, die Bearbei-

---

[45] Der Name Corproc steht für *corpus processing*. Für eine grobe Beschreibung s. Bartels 2020: Kap. 4.4.

[46] Alle in der Corproc-Umgebung durchzuführenden Prozesse, also auch die folgenden Schritte der Lemmatisierung und insbesondere der Behandlung der nicht automatisch lemmatisierten Tokens wurden von Richard Bígl und Božena Braumanowa vorgenommen, dasselbe gilt für die Erstbearbeitung der als „neue" Wortformen eingestuften Tokens für die Erweiterung der Lex-DB. Die Evaluierung und ggf. Korrektur dieses ersten Analyseschritts erfolgte durch Sonja Wölke.

[47] Die Webseite www.soblex.de „wird gemeinsam bereitgestellt durch das Sorbische Institut, die Stiftung für das sorbische Volk, das WITAJ-Sprachzentrum sowie Bernhard Baier und Wito Böhmak" (vgl. https://soblex.de/?cmd=about), die lexikografische Basis bildete zunächst das obersorbische orthografische Wörterbuch in der Ausgabe PS 2005. In den letzten Jahren wurde sie am SI im Rahmen von durch die Stiftung für das sorbische Volk geförderten

tung der nicht automatisch lemmatisierbaren Tokens. Hier ist im Hinblick auf das Ziel des Monitorings – die Identifizierung „neuer", d. h. nicht in der Lex-DB enthaltener Wörter und Wortformen – zunächst nur eine sehr grobe Klassifizierung möglich, einerseits in weiter zu bearbeitende Tokens (sog. Exportlösungen, drei Kategorien: in der Datenbasis fehlende Wortformen von Appellativa, Orts- und Personennamen einschließlich ihrer Ableitungen sowie Abkürzungen) und andererseits nicht für die Aufnahme in die Lex-DB vorgesehene Tokens (Ausschlusslösungen: Nicht-Wörter wie z. B. URLs, E-Mail-Adressen, aus Ziffern und Symbolen bestehende Tokens, Druckfehler, fremdsprachige Elemente und Passagen, alte und dialektale Wortformen), analog zu den in Kapitel 3.2.2.2 behandelten Erscheinungen. Gleichzeitig werden in diesem Schritt erforderliche Tokengrenzenkorrekturen markiert (z. B. Zusammenführung von Abkürzungen mit dem folgenden Punkt (|*med*|.| → |*med.*|; |*administr*|.| → |*administr.*| oder Trennung von Tokens wie in |*Drježdźany-Zhorjelc*| → |*Drježdźany*|-|*Zhorjelc*|). Die korrigierten Tokens können nach einem zweiten Durchlauf der automatischen Lemmatisierung dann weiterverarbeitet werden. Aus den Texten der hier berücksichtigten ersten drei Quartale des Jahrgangs 2019[48] wurden den Ausschlusslösungen 17 191 Tokens zugeordnet.

3.2.3.2 Ausschlusslösungen – für das Monitoring nicht oder gering relevantes Sprachmaterial

Entsprechend dem grundsätzlichen Ziel des Sprachmonitorings – der Identifizierung neuer, lexikografisch noch nicht erfasster Wörter bzw. Wortformen – stehen Ziffern, Symbole, E-Mail-Adressen, Initialen, Abkürzungen u. Ä. nicht im Fokus. Prinzipiell gilt dasselbe auch für Druckfehler (Buchstabendreher und -auslassungen, überflüssige Buchstaben) und orthografische Fehler, doch sind hier zum Teil systematische Beobachtungen zu machen.

Unter den orthografischen Fehlern fallen solche auf, die darauf beruhen, dass im Obersorbischen zum Teil gleich klingende Laute nach dem in synchroner Perspektive vom Laien nicht immer nachvollziehbaren etymologischen Prinzip durch unterschiedliche Grafeme wiedergegeben werden bzw. verschiedene Grafeme in der Aussprache ohne Entsprechung bleiben. So wird derselbe Laut [ʧʲ] im Obersorbischen durch <č> oder <ć> wiedergegeben, was zu Verwechslungen führt (*hotowarnićе* statt *hotowarniče*, *wot-*

Projekten um die sorbische Lexik aus dem Wörterbuch neuer Lexik (DOW-NL 2006), dem Terminologiemodul des neuen, in Arbeit befindlichen Deutsch-obersorbischen Wörterbuchs und weiteren Quellen (Mešk ank 2017, Rěčne kućiki, Exonyme) erweitert. Anfang 2021 erfolgte eine Erweiterung um zusätzliche Lexik aus dem DOW 1989/91, das Ende 2020 als Online-Version bereitgestellt wurde (www.obersorbisch.de/dow). Der für die Erzeugung der Flexionsformen verwendete morphologische Generator wurde 2016/17 im Sorbischen Institut revidiert und überarbeitet und wird in Abhängigkeit von der neu zu integrierenden Lexik laufend weiterentwickelt.

48 Gleichwohl konnten bis zum Verfassen des vorliegenden Berichts noch nicht alle obersorbischen Texte ausgewertet werden, daher stützen wir uns hier auf die Ergebnisse des Monitorings der ersten drei Quartale des Jahrgangs 2019. Angesichts der auszuwertenden Datenmenge für das Obersorbische halten wir die Aussagekraft der Auswertung durchaus für gegeben. Die Zahlenangaben (s. Tabelle 5) stellte Marek Slodička zusammen. Die Zunahme der Lemmatisierungsquote ist dadurch zu erklären, dass jeweils nach der Bearbeitung der Texte eines Quartals die als Datenbasis für Corproc dienende Lex-DB ergänzt und aktualisiert wurde.

paćenje statt *wotpačenje*, *zaklinći* statt *zaklinči*, *Miłoćanskej* statt *Miłočanskej*; *njepočeće* statt *njepoćeće*, *Wjelečanskeje* statt *Wjelećanskeje*). Ebenso werden mitunter <w> und <ł> verwechselt (beides gesprochen als [w]), wie in *zasyłaja* statt *zasywaja*, *płaćiły* statt *płaćiwy*, *zrawy*[49] statt *zrały*, dasselbe gilt für <k> und <ch> im Morphemanlaut (*kichota* statt *chichota*); <h> vor Konsonanten bleibt ohne Entsprechung in der Aussprache und fehlt daher mitunter im Schriftbild (*zromadźiznu* statt *zhromadźiznu*, *zładkowane* statt *zhładkowane*, hyperkorrektes *tuhdyšeje* statt *tudyšeje*). Zu ähnlichen Fehlern führt die Aussprache von <dn> zwischen Vokalen als [n] (*popołnju* statt *popołdnju*), die Aussprache von <błu> als [bu] (*kobłukom* statt *kłobukom*, hyperkorrektes *błudźaki* statt *budźaki*). Weiter wird <ě> vor <j> oft als [e] realisiert, worauf die nicht normgerechte Schreibung *póžreje* statt *póžrěje* zurückzuführen ist. Die Form *wobej* statt *wobě* ist dagegen durch Analogie zum Nom./Akk. Du. fem./neutr. der Adjektive zu erklären (z. B. *lubej*, dieselbe Analogiewirkung ist im katholischen Dialekt auch bei der Numeralform *dwej* statt *dwě* zu beobachten, vgl. SSA 10: 294 f.).

Zum Teil sind auch Änderungen der Orthografie, die erst in den letzten Jahren vorgenommen wurden,[50] in den Texten nicht berücksichtigt, z. B. *dioksida* statt *dioksyda*, *awdioguida* statt *awdijoguida*, *cityowej* statt *cityjowej*. Eine weitere Kategorie von orthografischen Regelverstößen, die für das Monitoring nicht relevante Wortformen erzeugen, sind nicht regelgerechte Adaptionen von fremdsprachigen Namen und Fremdwörtern wie *Šerjemjetjewo*, *Šeremjetjewo* statt *Šerjemjetewo* bzw. *resistentny* statt *rezistentny*, *inklusiwnje* statt *inkluziwnje*, *cenzora* statt *censora*.[51]

Als für das Monitoring nicht relevant wurden zudem (anders als beim Niedersorbischen, s. unter 3.2.2.3) zusammenfassende Schreibungen angesehen wie *(n)ostalgija* für *nostalgija resp. ostalgija*, *(dwu)rěčneho* für *dwurěčneho resp. rěčneho*; *tu- a wukraja* für *tukraja a wukraja*; *dźewjeć- do jědnaćelětnych* für *dźewjećlětnych do jědnaćelětnych*; nicht selten werden solche Schreibungen für genderneutrale/-gerechte Formulierungen eingesetzt wie in *pčołar(ka)* statt *pčołar abo pčołarka*, ähnlich *awtor/ka*, *informowana/y*, *rejwarjo/-ki*.

Probleme bereitet weiterhin die Tatsache, dass für das Obersorbische die Getrennt- bzw. Zusammenschreibung bis auf die Schreibung höherer Zahlenwerte bisher nicht systematisch und explizit geregelt ist,[52] sodass sich die Verfasser der Texte diesbezüglich nur auf den Usus stützen können sowie auf die Registrierung von Zusammenschreibungen in Wörterbüchern. Vor allem aufgrund der Zusammenschreibung der deutschen Äquivalente kommt es vor diesem Hintergrund nicht selten zur Zusammenschreibung sorbischer Mehrwortausdrücke (*mjezsobu* ‚untereinander' statt *mjez sobu*, *lětadołhej* ‚jahrelang' statt *lěta dołhej*, *přičimž* ‚wobei' statt *při čimž*). Auch diese Bildungen wurden zunächst aus der Menge potentieller „neuer" Wörter ausgeschlossen, zumal die Bestandteile schon in der Lex-DB registriert sind.

Auszuschließen waren darüber hinaus auch in den obersorbischen Texten fremdsprachliche Tokens, die vor allem als Bestandteile von Titeln (*Sächsische Zeitung*, *Asterix*

---

49 Möglicherweise unter dem Einfluss des verwandten Verbs *zrawić*.

50 Es handelt sich um Änderungen gemäß PS 2005 bzw. 2014 sowie HRK 2007.

51 Vgl. Adaptions- und Transkriptionsregeln in PS 1981: 599–608.

52 Entsprechende Regeln sind von der Obersorbischen Sprachkommission diskutiert und beschlossen, aber noch nicht veröffentlicht worden.

und das Geheimnis des Zaubertranks, *Sokol*[53] *v životě národa*, *Slavjanskoe jazykoznanie*), Namen von Firmen, Einrichtungen, Produkten u. Ä. begegnen (*Israel Aerospace Industries*, *Wendischer Freundes- und Arbeitskreis*, *Global Footprint Network*, *„Haus Bergland“*, *„Jägers Norddeutscher Champagnerroggen“*), aber auch, in Klammern gesetzt und manchmal mit der Sprachbezeichnung markiert, als erklärende deutsche Übersetzung zu sorbischen Appellativen – *při młynje na kózłach (Bockwindmühle)*, *kmanosć zapřijimanja (Auffassungsgabe)* – und Namen – *w Stróži (Wartha)*, *při Čornej Truze (Schwarze Lache)*, *Gojacki jězor (Schwielochsee)*. Während die so markierten Appellativa meist in der ersten Corproc-Bearbeitung als Ausschlusslösungen qualifiziert wurden (156 Tokens), geschah dies bei den Ortsnamen fast in der Hälfte der Fälle erst bei der späteren Bearbeitung – insgesamt wurden 473 solche in Klammern gesetzte erklärende deutsche Tokens gefunden.

Mitunter weisen aber auch bei fremdsprachigen Tokens, meist bei solchen mit Namencharakter, Flexionsformen darauf hin, dass hier Integrationsprozesse ablaufen, die sie als „sorabisierten“ Bestandteil des obersorbischen Textes qualifizieren, der somit im Fokus des Monitorings steht: *W septemberskim wudaću Česko-lužickeho věstnika* […], *w Bad Lauterbergu*, *Populistiska alianca prawicarskeje Legi a hibanja Pjeć hwězdow* […].

Auch außerhalb des Systems der obersorbischen Standardsprache der Gegenwart stehende sorbische Tokens werden ausgeschlossen, sie werden bei der Bearbeitung der nicht automatisch lemmatisierten Tokens in Corproc als „isOld“ bzw. „isDialect“ markiert. Das betrifft vor allem zitierte Textpassagen bzw. Wörter, wie z. B. im bearbeiteten Jahrgang in einer namenkundlichen Artikelserie bzw. in geschichtswissenschaftlichen Beiträgen in der Zeitschrift „Rozhlad“, ebenso in verschiedenen Buchpublikationen, wie in Os-3, einer Anthologie mit historischen Zeitungsbeiträgen, deren Sprache nicht vollständig der aktuellen Norm angepasst wurde. Schwierig wird die Abgrenzung mitunter in Os-17 (dort Cušcyna), wo die Verflechtung ober- und niedersorbischen sowie historischen oder dialektalen Sprachmaterials als künstlerisches Mittel eingesetzt wird.

3.2.3.3 Weitere Bearbeitung der qualifizierten Tokens

Die als mögliche Erweiterungen der Lex-DB eingestuften Tokens (in Corproc sog. Exportlösungen der Typen Wortformen (missLex), Namen (isName) oder Abkürzungen (isAbbr) wurden einschließlich Kontext, Quellenangaben und IDs quartalsweise zusammengefasst und über Zwischenschritte jeweils in die Lex-DB eingearbeitet. Das geschah für das erste und einen Teil des zweiten Quartals durch direkte Integrierung in die Datenbasis von soblex einschließlich der Zuordnung der entsprechenden Flexionsmuster und mitunter notwendiger Modifikationen des morphologischen Generators (insgesamt 6803 Lexeme, davon 4318 Personen- oder Ortsnamen sowie 2485 Appellativa und Namen anderer Kategorien). Da sich dies aber als zu zeitaufwendig erwies, wurden später nur die tatsächlich in den ausgewerteten Texten vorkommenden Formen jeweils mit zugeordneter Grundform aufgenommen.[54] Entsprechend steigerte sich wie erwartet die Quote der

---

[53] Hier ein sorbischer Eigenname in einer tschechischen Überschrift.

[54] Die vollständige Integration in die Lex-DB, einschließlich Flexionsformen, ist ein weiterer Schritt, der erst später, nach genauerer Beurteilung und Qualitätskontrolle erfolgen soll. Auch

automatischen Lemmatisierung – zunächst nach der Integration der gefundenen neuen Lexik in den morphologischen Generator recht schnell (3,07 Prozentpunkte), später etwas langsamer (0,92 Prozentpunkte):

Quartal	2019/I	2019/II	2019/III	I–III insg.
Dateien	22	77	88	187
Tokens	152 076	510 871	603 834	1 266 781
davon unlemmatisiert	10 973	21 220	19 500	51 693
Lemmatisierungsquote[55]	92,78 %	95,85 %	96,77 %	95,92 %

Tab. 5: Ergebnisse der automatischen Tokenisierung und Lemmatisierung der obersorbischen Texte mittels Corproc

Bei der Bearbeitung zeigte sich, dass trotz detaillierter Instruktionen auch bei der ersten groben Klassifizierung des nicht automatisch erkannten/lemmatisierten Sprachmaterials unter Corproc Subjektivität bei der Beurteilung nicht auszuschließen ist. Daher finden sich auch unter den Exportlösungen noch Schreibfehler und nicht integrierte systemfremde Tokens (fremdsprachlich, älteren Sprachschichten zugehörig, dialektal) und auch Tokens, die als Nicht-Wörter anzusehen sind, s. o., die dann erst in der zweiten Bearbeitungsrunde entsprechend gekennzeichnet werden.

Insgesamt wurden 33 705 Tokens als Exportlösung eingestuft (18 072 unterschiedliche), davon wurden nachträglich 1142 als Tokens mit Nicht-Wort-Charakter, fremdsprachig, Druckfehler, Zitate aus älteren Sprachzuständen aussortiert, somit verblieben 16 930 unterschiedliche Tokens. Letzteren wurden jeweils die Wortgrundformen zugeordnet (14 176 unterschiedliche Grundformen, berechnet per Pivot-Tabelle). Den Löwenanteil der Exportlösungen machen Tokens aus, die Eigennamen oder deren Bestandteile sind (vgl. 4.1). Im Endeffekt ergeben sich aufgrund der nicht automatisch erkannten Tokens abzüglich der Orts- und Personennamen und ihrer Derivate 5783 unterschiedliche Wortgrundformen, die jedoch weiter zu beurteilen sind, inwieweit sie als Ergänzung der Lex-DB in Frage kommen. Als besonderer Umstand in dieser ersten Phase des Monitorings war zu berücksichtigen, dass davon 829 (14,4 %) schon im DOW 1989/91 gebucht sind, dieses Wörterbuch aber noch nicht in die obersorbische Lex-DB integriert war.[56] Zum Teil handelt es sich dabei um Wortformen, die aus heutiger Sicht veraltet sind (im DOW 1989/91 oft mit Qualifikator *veraltend* bzw. *veraltet* versehen, z. B. *kasarna* statt *kaserna* bzw. *přecy* statt *přeco*).

---

die schon für den morphologischen Generator bearbeiteten neuen Lexeme müssen vor der endgültigen Integration noch einer genaueren Beurteilung unterzogen werden, inwieweit sie zur schriftsprachlichen Norm gehören.

55 Die hier für den gesamten Zeitraum angegebene Lemmatisierungsquote von 95,92 % (vgl. Fußnote 33 zu Tab. 3) liegt sehr nah bei der im Niedersorbischen erreichten von 96,25 %. Damit konnten schon im ersten Monitoringjahr aufgrund der Vorarbeiten (s. Bartels 2020) sehr gute Ergebnisse erzielt werden.

56 Durch die Ergänzung eines Teilbestands der Lexik aus dieser Quelle Anfang 2021 kann die Anzahl der automatisch erkannten Tokens zusätzlich erhöht werden (s. Fußnote 28).

4. Ausgewählte Ergebnisse

Im Folgenden werden Ergebnisse der auf diese Weise erstmals durchgeführten Datenanalyse für das verfügbare Schrifttum 2019 vorgestellt. Dabei werden weitere Elemente des noch in der Entwicklung befindlichen Verfahrens wie auch die Vielfalt der beobachteten Phänomene deutlich.

Kap. 4.1 verweist zunächst erneut auf den besonderen Status von Eigennamen, die in ihrer großen Anzahl vor allem in der hier beschriebenen Frühphase des Monitoring-Vorhabens für das gesamte Analyseverfahren eine besondere Herausforderung darstellen. Obgleich Registrierung, Beschreibung und Darstellung im Wesentlichen in anderem Rahmen erfolgen wird, spielen Eigennamen zumindest am Rande auch in den weiteren Abschnitten dieses Kapitels eine Rolle.

Kap. 4.2 thematisiert relativ häufig auftretende „Abweichungen" verschiedener Art, die bei den weiter zu analysierenden Textwörtern vorkommen und eine Bewertung erfordern, die zu formaler Anpassung, Aussonderung oder Einbeziehung in die Folgeanalyse führt.

Kap. 4.3 präsentiert dann bereits um fehlerhafte Formen weitgehend bereinigte „neue Lexik" im Sinne der Erläuterungen aus Kap. 2.2.

Im vorletzten Abschnitt 4.4 gehen wir kurz auf den Nutzen des Schrifttumsmonitorings für Sprachbeschreibung, Sprachkodifizierung und Sprachunterricht ein. Abschließend wird – mit Blick auf entsprechende Erwartungen – in Kap. 4.5 erläutert, warum der derzeitige Entwicklungsstand elaborierte sprachstatistische Auswertungen noch nicht zulässt.

4.1 Eigennamen

Wie bereits an verschiedenen Stellen erkennbar, kommt Eigennamen als Teil der nicht erkannten Lexik – quantitativ wie auch bzgl. ihrer „inhaltlichen" Relevanz – eine besondere Bedeutung zu. Texte handeln von Dingen, die vielfach mit Eigennamen bezeichnet werden, und zwar unabhängig davon, ob es sich um „sorbische" Namen handelt oder nicht. Eine Unterscheidung zwischen „sorbischen" und „nicht sorbischen" Namen ist vielfach problematisch. Abgesehen davon sind alle Namen, die in den sorbischen Texten auftreten, als „Wörter" gleichermaßen relevant, weshalb es nicht sinnvoll ist, sie ähnlich wie „nicht-sorbische Lexik" auszusortieren. Eigennamen in sorbischen Texten – auch „nicht-sorbische" – sind häufig morphologisch adaptiert und liegen in der Regel in flektierter Gestalt vor.

In den niedersorbischen bzw. obersorbischen Lex-DB, die der automatischen Textanalyse im ersten Monitoring-Jahr zugrunde lagen, waren Eigennamen nur insoweit erfasst, als sie in den für die Erstellung der Lex-DB ausgewerteten Wörterbüchern und ergänzenden Quellen verzeichnet waren.[57] Es war daher zu erwarten, dass „unerkannte" Textwörter des Typs „Eigenname" in großer Menge auftreten würden.

---

[57] Für das Niedersorbische waren das im Wesentlichen: DNW (2003 ff.), Starosta 1999 und Mucke (1911–28). Aus diesen Quellen wurden insgesamt 9200 Eigennamen (Lexeme) in die ns. Lex-DB aufgenommen. Für das Obersorbische gibt es zwei Quellen, die ausschließlich Eigennamen enthalten (Meškank 2017, Exonyme), auch in weiteren Quellen der von Soblex generierten obersorbischen Lex-DB (PS 2005 bzw. 2014, DOW-NL 2006, Terminologiemodul des neuen DOW) sind Eigennamen enthalten, und zwar neben Personen- und Ortsnamen auch weitere Namenkategorien wie Länder-, Völker-, Gewässernamen u. a. Insgesamt enthielt die Ausgangsversion der os. Lex-DB (vor dem Monitoring) ca. 5600 unterschiedliche Eigennamen einschließlich ihrer Flexionsformen.

Für das Niedersorbische machen die Eigennamen etwa 65 % der nicht erkannten Lexeme aus (n = 4257). Für das Obersorbische machen Personen- und Ortsnamen in den ausgewerteten drei Quartalen 59 % (n = 8468) der nicht erkannten Wörter aus, weitere 13 % (n = 1809) sind Eigennamen anderer Kategorien (Namen weiterer topografischer Objekte, Organisations- und Institutionsnamen u. Ä.).

Dabei bilden Vornamen (*Jörg*, *Beata*) und Familiennamen (*Brězanowa*, *Kosacojc*, *Njepilic [dwór]*) sowie adjektivische (possessivische) Ableitungen von ihnen (*Erwinowy, Janašowy*) sowohl im Niedersorbischen (51 %, n = 3292) als auch im Obersorbischen die größten Gruppen.

Eine weitere umfangreiche Gruppe (ns. 9 %, n = 587; os. 11 %, n = 1543) bilden substantivische und adjektivische Toponyme, vor allem Ortsnamen unterschiedlicher Art (ns. *Chróśćicy*, *Essen*, *nowoniwjański*, *hannoverski*, os. *Brunšwik*, *Błobošojce*, *Auerbachski*), Namen von Ländern, Regionen usw. (ns. *Tyrolska*, *kroatiski*, os. *Wojwodina*, *Podkarpatska*) sowie diverser anderer topografischer Objekte (ns. *Karkonoše*, *srjejźomórski*, os. *Satkula*, *Šmrěčnikowy*). Daneben traten natürlich auch andere Namentypen auf: Bezeichnungen für Personengruppen (ns. *Dolnonimc*, *łemkojski*, os. *Bělochorwat*, *Głomačan*), Sprachen (ns. *fering*, *sanskritski*, os. *okcitanšćina*, *quechua*), Institutionen (ns. *rbb*, *maćicny*, os. *Leopoldina*, *ZVON*, *MWFK*), Marken (ns. *Focke-Wulf*, os. *Microsoft*), Titel (ns./os. *Europeada*, os. *Dźěćiznak*) und diverse andere (ns. *Ostwall*, *Bimmelgusta*, os. *PowerSerb*, *Matješk-PowerTools*).

Die im ersten Monitoring-Jahr ermittelten „neuen“ Formen von Eigennamen dienen als ergänzende Datengrundlage für das Monitoring der Folgejahre und werden so schrittweise die Menge jeweils unregistrierter Namensformen reduzieren. In Abhängigkeit von bestimmten Kriterien (Vorkommenshäufigkeit u. a.) werden sie darüber hinaus in weitere Sprachressourcen integriert: in die regulären Lex-DB, in die Datenbasen für die automatische Rechtschreibkontrolle sowie in die Internetseite „Sorbische Namen“.[58] Eine Einbindung neuer Eigennamen in diese Ressourcen erfordert – in steigendem Maße nach der Reihenfolge der Nennung – eine teils intensive Nachbearbeitung.[59] Eine begründete Auswahl ist daher notwendig.

Der auf diese Weise organisierte Aufbau einer zunehmend umfassenden Datenbasis aller im sorbischen Schrifttum belegten Eigennamen wird in Zukunft ein wichtiges Instrument zur zuverlässigen inhaltlichen Erschließung der digitalisierten Korpustexte sein: Über den Weg der Identifizierung und Annotation zunächst einzelner Namensformen sowie später, darauf aufbauend, auch mehrteiliger und mehrsprachig gebräuchlicher Namen (z. B. von Personen oder Institutionen: *Arnošt Muka* / *Ernst Mucke*, *Załožba za serbski lud* / *Stiftung für das sorbische Volk*) wird ein wichtiger Zugang zum digitalisierten Schrifttum geschaffen (s. Bartels 2020, insb. Kap. 4.4.4).

---

[58] Da derzeit noch auf niedersorbische Eigennamen beschränkt, zugänglich ausschließlich über das Sprachportal www.niedersorbisch.de/mjenja. Eine Erweiterung um obersorbische Namen soll nach Möglichkeit in weiteren Projektphasen erfolgen. Das Vorhaben wurde nach zweijähriger Förderung Ende 2020 vorerst beendet.

[59] So ist zum Beispiel für eine Integration in die Datenbasis der automatischen Rechtschreibprüfung eine abschließende Klärung der Normgerechtigkeit sowie der Flexionsweise notwendig. Auf dieser Grundlage werden dann alle auch nicht in den Korpustexten belegten Flexionsformen gebildet. Eine besondere Herausforderung bilden hier Doppelnamen, deren Bestandteile häufig separat und unterschiedlich flektiert werden (z. B. os. *Jěwa-Marja*, *Bart-Ćišinski*, *Pančicy-Kukow*).

4.2 Nicht-standardsprachliche Elemente

Unter den Formen, die durch das derzeitige, oben beschriebene Verfahren nicht erkannt werden, findet sich eine Reihe von Belegen, die nicht der Standardsprache zugeschrieben werden können, sondern eher in einem nicht immer klar abgrenzbaren Kontinuum zwischen dialektalen und umgangssprachlichen Formen angesiedelt sind. Es handelt sich dabei oft um (lautliche oder morphologische) Abweichungen von registrierten standardsprachlichen Formen. In der Mehrzahl der Fälle werden die im Folgenden vorgestellten Phänomene als nicht normgerecht aussortiert, sie sind also bspw. für die Rechtschreibkontrolle bzw. für den lexikalischen Ausbau der Standardsprache nicht relevant. In Einzelfällen wird auf einen eventuellen Änderungsbedarf der geltenden Kodifizierung hingewiesen.

4.2.1 Niedersorbisch

4.2.1.1 Lautliche Abweichungen

4.2.1.1.1 Vokale

Meist eher dialektal einzuordnen sind Formen, in denen ein Vokal der Standardsprache durch einen anderen ersetzt wurde. Dieser Typ tritt nicht häufig auf und ist vielfach Bestandteil von Zitaten aus älteren Texten:

- *o* statt *e*: *trjobny* statt *trjebny*, *wjasołšy* statt *wjaselšy*
- *o* statt *a*: *kolendaŕ* statt *kalendaŕ*, *połkanje* statt *pałkanje*
- *u* statt *y*: *budlenje* statt *bydlenje*, *buwaś* statt *bywaś*, *pśebuwalnik* statt *pśebywalnik*
- *u* statt *o*: *šulta* statt *šołta*
- *e* statt *(')a*: *kermuša* statt *kjarmuša*, *pśeklepany* statt *pśeklapany*[60]
- *a* statt *e*: *parika* statt *perika*
- *'e* statt *ě*: *prjeny* statt *prědny*

Eher nicht dialektal bzw. umgangssprachlich sind die Neuentlehnungen (vgl. 4.3.2.4) *tafla*, *taflicka* statt *tofla*, *toflicka*. Sie könnten als Phänomene, die mit der Sprachverwendung durch Neusprecher verbunden sind, aber auch als Obersorabismen gewertet werden.

Umgangssprachliches Auslassen des Vokals spiegelt sich wider in *intresantny*, *intresěrowaś*. In diesem Fall ist bereits die Nutzung von *int(e)res* eher umgangssprachlich, gegenüber in der Standardsprache präferiertem *zajm*. Das fehlende *i* im zweimal belegten *wjelki* statt *wjeliki* ist dagegen höchstwahrscheinlich als Lapsus[61] zu werten.

Mehrfach mit Ausfall von *a* belegtes dialektales *kótryž* statt standardsprachlichem *kótaryž* ist hauptsächlich (aber nicht ausnahmslos) in Zitaten aus älteren Quellen zu finden.

---

[60] Mit *e* statt *a* ist außerdem als Zitat aus dem Schleifer Dialekt, in dem der entsprechende Lautwandel generell nicht eingetreten ist, belegt: *kerluš* (‚kjarliž').

[61] Je nach Autor könnte es sich z. B. um einen Polonismus handeln.

4.2.1.1.2 Konsonanten

Auf dialektaler Basis steht *mj* statt *n* in *mjerk* (statt *nerk*), *mjerkanje* (statt *nerkanje*). Interessanterweise ist das kodifizierte *nerk* im Material des Sorbischen Sprachatlas (SSA 3: 132 f.) gar nicht belegt, es treten hauptsächlich Formen vom Typ *mjerk* und *jerk*/*jark* auf. Die einzige mit *n*-Prothese anlautende Form (Typ *njark*) ist für Schmogrow belegt. Für dieses Lexem wäre eine Änderung der aktuellen Norm vorstellbar. Ebenfalls *m* statt des kodifizierten *n* steht in *brošma* (statt *brošna* bzw. *brošno*).[62]

Typisch für die meisten Dialekte ist die durch Dissimilation entstandene Form *cašny* statt *tšašny*. Das Auftreten dieser Aussprachevariante in gedruckten Texten ist wiederum auf Zitate aus der älteren Literatur zurückzuführen.

Die Schreibungen von *w* statt *j* in *pśestawiś* statt *pśestajiś*, *pśestawjony* statt *pśestajony*, *pśistawiś* statt *pśistajiś*, *stawiś* statt *stajiś*, *wustawiś* statt *wustajiś*, *wustawjony* statt *wustajony*, *zadawiś* statt *zadajiś*, *zatawjony* statt *zatajony*[63], *zawóstawiś* statt *zawóstajiś* sind (zumindest mit Blick auf die heutigen Verhältnisse) nicht als dialektaler Einfluss zu werten. Es handelt sich zumeist um historische Schreibungen in Zitaten und in einigen Fällen möglicherweise um bewusste Historisierung.[64]

Für viele Dialekte (und zum Teil für die Umgangssprache) typisch ist die Vereinfachung bestimmter Konsonantengruppen. Dies schlägt sich in den heutigen Drucken kaum nieder. In den untersuchten Dokumenten finden sich wenige relevante Beispiele, meist im Kontext von Zitaten aus der älteren, weniger stark normierten bzw. in der Schreibung mehr an die Aussprache angelehnten Literatur.

- *n* statt *dn*: *krynjenje* statt *krydnjenje*, *krynuś* statt *krydnuś*, *zabynuś* statt *zabydnuś*, *pŕěny* bzw. *prjeny* statt *pŕědny*
- *n* statt *gn*: *pózwinuś* statt *pózwignuś*, *pśeternuś* statt *pśetergnuś*, *śěnuś* statt *śěgnuś*
- *rs* statt *rbs*: *serski* statt *serbski*

4.2.1.1.3 Schnellsprechformen

Für die Umgangssprache und die Dialekte typisch sind weitere Vereinfachungen bzw. Schnellsprechformen. Dazu gehören *bdu/b'du* statt *budu*, *bźo* statt *buźo*, *bźomej* statt *buźomej*, *bźomy/b'źomy* statt *buźomy*, *njeb'źoš* statt *njebuźoš*, *njebdu/njeb'du* statt *njebudu*, *njebźo* statt *njebuźo*, *njebźomy* statt *njebuźomy*, dazu *ja b'* statt *ja budu*; *njamgu* statt *njamógu*, *njamžo/njam'žo* statt *njamóžo*, *njamžoš/njam'žoš* statt *njamóžoš*; *njej'* statt *njejo*. Diese und weitere Formen finden zwar in den grammatischen Tabellen der dem Monitoring zugrunde liegenden Wörterbücher – und damit auch bislang in der Lex-DB – keine Erwähnung, nichtsdestotrotz treten sie in einigen Wörterbuchbeispielsätzen auf (zusätzlich zu den genannten z. B. *njeboš* statt *njebuźoš*).

---

62 Bei Mucke (1911–28) ist die Zuordung von dialektaler und standardsprachlicher Form umgekehrt zu den aktuellen lexikografischen Beschreibungen.

63 Während in den anderen Beispielen heutiges *j* auf *ẇ* zurückgeht, ist die Schreibung mit *wj* (und historisch wohl auch die phonetische Realisierung als [wʲ]) in *zatawjony* hyperkorrekt, da das zugrunde liegende Verb urslawisch **tajiti* kontinuiert.

64 Derartige Belege wurden zum Teil bereits zuvor aussortiert, vgl. Kap. 3.2.2.2.3.3.

4.2.1.2 Abweichungen in Flexion und Wortbildung

Auch im Bereich der Konjugation und Deklination sind in den untersuchten Texten Formen belegt, die in der aktuellen Kodifikation keine Berücksichtigung finden. Dies betrifft u. a. die Form *(ja) debu* statt *dejm*, *(wóni) coju* statt *kšě*. Auffällig ist die bisher nicht belegte Dualform *gólcyma* statt *gólcoma*.[65]

Relativ oft belegt ist die Endung *-oju* statt *-eju* bei personenbezeichnenden Substantiven auf *-ař*: *fararjoju*, *góspodarjoju*, *guslowarjoju*, *kucharjoju*, *mejstarjoju*, *ministarjoju*, *reportarjoju*, *spiwarjoju*. Die meisten Belege stammen jedoch aus einer einzigen Quelle, der (überarbeiteten) Neuauflage der Krabat-Erzählung von Měrćin Nowak-Njechorński in der Übersetzung des aus Drewitz gebürtigen Wylem Bjero (Ns-6). Während im Material für den Sorbischen Sprachatlas offenbar kein Beleg dieses Typs auftritt (vgl. SSA 11: 140), gibt es in älteren Schriften hunderte Beispiele dafür.

Nach derzeitigem Stand der Kodifikation ist *-ejski* statt *-ojski* in *smogorjejski* als dialektal zu bewerten. Auffällig ist jedoch, dass von 29 relevanten Stellen im Korpus des muttersprachlichen Niedersorbisch (https://dolnoserbski.de/dobes) höchstens ein Beleg (falls überhaupt) als Lautung mit *-o-* gewertet werden könnte. Da für den Ort *Kokrjow* die Adjektivform *kokrjejski* kodifiziert ist, wäre eine entsprechende Änderung der Norm *smogorjojski* zu *smogorjejski* zu erwägen.

Die häufigen Belege für den Gen. Pl. von Maskulina mit weichem Stammauslaut auf *-i* (Typ *cantnari*) sind hingegen als standardsprachlich zu werten. Die Endung findet u. a. bei Janaš (1984: 114) generell als fakultative Variante und vorsichtiger formuliert bei Starosta (1992: 45) Erwähnung.[66] Dass die entsprechenden Formen zurzeit beim Monitoring als „nicht registriert“ gewertet werden, resultiert aus der bisher vereinfachenden Beschreibung in den Wörterbüchern, die die Datenbasis für die notwendigen Erkennungsprozesse bilden.

In der Standardsprache wird *l* generell als phonologisch weicher Laut gewertet. Bei Muttersprachlern ist er – mit Konsequenzen für die Flexion – in verschiedenen Fremdworttypen sowohl phonetisch als auch phonologisch hart. Auch hierfür finden sich im untersuchten Material Belege, beispielsweise Nom. bzw. Akk. Pl. (nach den Deklinationsmustern von Wörtern mit hartem Stammauslaut) auf *-ly* statt *-le*: *akwarely*, *cily*. Hierzu gehört wohl auch der Beleg *roly* (zu: *rola* ‚Rolle im Theater‘[67]). Aus anderen Gründen gibt es bei einigen Neusprechern generell die Tendenz, die phonologische Weichheit von *l* aufzugeben, was die Schreibung *kraly* statt *krale* dokumentiert.

Die in den Dialekten mit unterschiedlicher Frequenz (vgl. SSA 12: 14–22) verbreitete Kürzung der Genitivendungen der Adjektive von *-ego* (*-ogo*) zu *-eg* (*-og*), die in der Standardsprache nicht ausdrücklich als normwidrig abgelehnt wird und auch in Beispielsätzen des DNW zu finden ist, spiegelt sich in den untersuchten Texten in einer Vielzahl von

---

[65] Der Plural vom Typ *gólcymi* (standardsprachlich *gólcami*) hingegen ist in historischen Texten wie auch im Korpus des muttersprachlichen Niedersorbischen (https://dolnoserbski.de/dobes) gut belegt. Dualformen auf *-yma*/*-ima* treten in den Dialekten auch jenseits der standardsprachlichen *wócyma*, *wušyma* auf.

[66] Die Nichterwähnung von *-i* in diesem Kontext in den neuesten grammatischen Tabellen (Šrejdař/Zakař 2017) ist daher eher kritisch zu bewerten.

[67] Das *l* in *rola* (‚Ackerland‘) und *rula* (‚Rolle (Walzenförmiges)‘) hingegen ist bei Muttersprachlern phonologisch grundsätzlich weich.

Belegen wider. Dabei handelt es sich sowohl um historische Zitate, als auch um aktuelle Beispiele.

Mit Blick auf die Wortbildung fallen einige Komposita ins Auge, deren erster Teil eine ursprüngliche Genitivform darstellt. Solche gibt es zwar auch in der (aktuellen) Standardsprache, sie sind jedoch selten, z. B. *głowybólenje*.[68] Bei den relevanten Belegen in den untersuchten Texten handelt es sich teilweise um Zitate aus der älteren Literatur, teilweise um eine bewusste Stilisierung. In den Dialekten sind solche Bildungen möglich. Es scheint jedoch, dass sie auch hier – wie in der Standardsprache – in vielen Fällen durch Formationen mit Akkusativ im Erstglied ersetzt werden bzw. ersetzt worden sind. Belege mit Gen. Pl. sind: *dušowpastyŕ*, *dušowpastyŕka*, *dušowpastyŕski*, *dušowpastyŕstwo*, *górkow-wopytowanje*, *idejow-dawaŕ*, *kulkow-kopanje*, *kulkow-kubłanje*, *rybow-łojenje*, *rybowłojśe* bzw. *rybow-łojśe*, *seśow-źaśe*, *zubow-bólenje*. Der Gen. Sg. ist belegt im ersten Teil von *głowy-wótryśe*, *klěba-pjacenje*, *stoga-kłaźenje*, *trajdy-mlaśe*. Im Prinzip handelt es sich um eine Inversion und Univerbierung der weit verbreiteten Genitivkonstruktionen (*bólenje zubow*, *kopanje kulkow*), worauf auch die fast ausnahmslose Schreibung mit Bindestrich hindeutet. Den oben genannten Beispielen entsprechen in den aktuellen Wörterbüchern entweder Formen mit Akkusativ im Erstglied (*dušepastyŕski*,[69] *kulkikopanje*, *rybyłojenje*, *zubybólenje*; *klěbpjacenje*), oder es könnten solche ohne weiteres gebildet werden (z. B. *kulkikubłanje*).

Lehnprägungen vom Typ *jazzmuzika* ‚Jazzmusik' sind auch im standardsprachlichen Niedersorbischen keine Seltenheit. Hier besteht jedoch eine weit größere Tendenz, stattdessen eine Konstruktion *Adjektiv + Substantiv* zu benutzen, vgl. *hobbyjowa piwnica* ‚Hobbykeller'. Lehnprägungen vom ersten Typ, die in der Lex-DB nicht erfasst waren, sind: *folk-party*,[70] *folk-projekt*, *grafikdesignaŕka*, *hobby-filmaŕ*, *hobby-fotograf*, *hobby-gerc*, *inwestigatiw-reportaŕ*, *kultusministaŕ*, *pancerjeep*, *party-śpa*, *pěś-procentklawsula*, *pilot-projekt*, *postagentura*, *radio-wusćełanje*, *rajchsministaŕ*, *regionalpolitiski*, *reprint-wudaśe*.

4.2.1.3 Nicht-standardsprachliche Lexik

Um typisch dialektale Lexik handelt es sich bei *brjaschen* statt *rjaschen*, *butranica*[71] statt *buternica*, *jermank* statt *mark*, *kachlink* statt *kachleńk*,[72] *klaška* statt *gła*, *kołaŕnja* statt *kołoźejaŕnja*, *kracchen* statt *rjaschen*, *kužoł* statt *kružoł*,[73] *lytyšk* statt *łucywko*, *pjerźowaty*

---

[68] Hier ist *głowy* als Gen. Sg. zu interpretieren, vgl. *bólenje głowy*. In *zubybólenje* jedoch ist *zuby* als Akk. Pl. zu betrachten.

[69] Die Form des Erstglieds *duše* könnte allerdings auch als Gen. Sg. gewertet werden.

[70] Im Original tritt fast ausschließlich die Schreibung mit Bindestrich auf. In der Standardsprache wird eine Schreibweise ohne denselben präferiert.

[71] Möglicherweise statt *butrjanica* (vgl. Mucke 1911–1928: *butŕanica*).

[72] Die Zuordnung *kachleńk* – standardsprachlich und *kachlink* – dialektal ist wohl eher zufällig, da die zweite Variante mindestens genauso verbreitet ist wie die erste (vgl. SSA 10: 94 f.). Sie könnte aus der Darstellung in Mucke 1911–1928 resultieren, wo unter dem in alphabetischer Reihenfolge weiter vorne stehendem Stichwort *kachleńk* ohne stilistische Unterscheidung beide Varianten beschrieben werden, wohingegen unter dem alphabetisch später einsortierten Stichwort *kachlink* lediglich ein Verweisartikel auf *kachleńk* steht.

[73] Im konkreten Beleg könnte es sich allerdings auch um einen Obersorabismus handeln: Das Beispiel stammt aus einer entsprechenden Übersetzung.

statt *pjerźaty*,[74] *pług* statt *chołuj*, *schajźaś* statt *schadaś*, *smargula* statt *smarž*, *šmokowaś* statt *słožeś*,[75] *tencajšny* statt *tencasny*, *wada* statt *wata*. Im Fall *šora* statt *slě* kann aufgrund fehlender Informationen nicht bestimmt werden, ob es sich bei *šora* um eine dialektale oder ältere überregionale Entlehnung aus dem Deutschen (zu *Geschirr*) handelt.[76]

Bei der umgangssprachlichen Lexik fällt vor allem die extensive Verwendung (älterer und neuerer) Entlehnungen aus dem Deutschen auf, bis hin zur auch im Schriftbild nicht adaptierten Verwendung deutscher Wörter. Auch hier gehen einige, jedoch bei weitem nicht alle Belege auf Zitate aus älteren Quellen zurück. In einigen Fällen wirkt sich das Fehlen eines passenden standardsprachlichen Äquivalents aus.

Einige konkrete Belege sind: *bonhof* statt *dwórnišćo*,[77] *cajtunga* statt *casnik*, *elektrańja* statt *milinańja*, *enkelgóle* statt *źišiźiše*, *erdbeerka* statt *słynicka*, *fichtny* statt *mokšowaty* bzw. *włožny*, *flejšař* statt *rěznik*, *forwark* statt *wudwór*, *frejda* statt *wjasele* bzw. *radosć*, *gešeft*[78] statt *not*, *knajpa* statt *kjarcma*, *landpartija* (bisher ohne lexikografisch registrierte standardsprachliche Bezeichnung), *lazowańica* statt *cytańka*, *opa* statt *starki*, *ortgang* (bisher ohne lexikografisch registrierte standardsprachliche Bezeichnung), *regierunga* (wohl besser zu schreiben als *regěrunga*) statt *kněžarstwo*, *rejza* statt *drogowanje*, *strejtarski/štrejtarski*[79] statt *rozestajeński*, *štempelowy*[80] statt *kołkowy*[81], *tabelka*[82] statt *tabulka*, *tuderski* statt *tudejšy*[83] *wyrtšaft* statt *žywnosć*[84]. (Eine

---

[74] Die Form *pjerźaty* ist in den Wörterbüchern nicht verzeichnet. Sie ist aber zu *pjerźeś* bildbar, wohingegen die Grundlage für *pjerźowaty* dialektales *pjerźowaś* ist.

[75] Die Wörterbücher verzeichnen als umgangssprachliche Variante *šmekowaś*, das eher im Westen des niedersorbischen Sprachgebiets verbreitet ist, wohingegen *šmokowaś* im gesamten Osten benutzt wird. Das standardsprachliche, auch im älteren Schrifttum verbreitete *słožeś* mit potenziellen Varianten (z. B. *słoźiś*) ist hingegen im Material des Sorbischen Sprachatlas (vgl. SSA 10: 138 f.) kaum belegt.

[76] Eine Aufzählung der Dialektwörter wird im Internet als Teil einer Datei bereitgestellt (s. dort Abschnitt K-23). Zur erwähnten Datei vgl. den Beginn von Kap. 4.3.2.

[77] Bei der Schreibung *bonhof* scheint es sich um eine bewusste Stilisierung zu handeln. Die Lautung von *bon-* mit *-o-* ist wohl dem nicht-standardsprachlichen sorbischen Äquivalent von deutsch *Bahn* (*bona*) entnommen. Interessanterweise verwenden sämtliche Muttersprachler im Korpus des muttersprachlichen Niedersorbischen (https://dolnoserbski.de/dobes) ausschließlich die Lautung mit (nicht langem) *a*: *banof*. Die einzigen Belege mit *o* stammen von Exploratoren.

[78] Im konkreten Beleg steht das Wort im Äquivalent der Wendung *sein Geschäft verrichten*.

[79] Beide Varianten sind als Teil des Mehrwortlexems *strejtańske/štrejtańske rozgrono* ‚Streitgespräch' belegt, wofür in den Wörterbüchern *rozestajeńske rozgrono* und *rozestajenje* zu finden ist.

[80] Besser wäre wohl *štemplowy*, vgl. in den Wörterbüchern verzeichnetes, als umgangssprachlich markiertes *štemplowaś*, *wótštemplowaś*.

[81] Die standardsprachliche Bedeutung ‚Gerät zum Stempeln' für *kołk* scheint ein Obersorabismus zu sein.

[82] In den Wörterbüchern findet sich neben *tabula* und dem zugehörigen Deminutiv *tabulka* auch (ohne stilistische Wertung) *tabela*. Streng genommen ist also *tabelka* als regulär gebildete Deminutivform zu *tabela* als standardsprachlich zu bewerten.

[83] *Tuderski* ist eine der wenigen belegten umgangssprachlichen Formen, die keine Entlehnung aus dem Deutschen darstellen. Die aktuell kodifizierte Form tritt im älteren Schrifttum nicht auf, wohingegen *tuderski* (in verschiedenen Schreibungen) wohl die meistverbreitete Entsprechung für deutsches *hiesig* darstellt.

[84] Als Äquivalent für (schon im Deutschen umgangssprachliches) *Klitsche* (ärmliches Anwesen) ist bereits *wirtšaftchenk* registriert, das wohl besser *wyrtšaftcheńk* geschrieben werden sollte.

vollständige Aufzählung der hier relevanten Fälle wird ebenfalls in einer gesonderten Datei im Internet bereitgestellt.[85])

Auf nicht-standardsprachlicher Aussprache basiert maskulines *kinow* gegenüber dem standardsprachlichen Neutrum *kino*.

Typisch für den Nonstandard ist die nicht adaptierende Entlehnung einiger Adjektive, die dann indeklinabel sind (Typ *fajn*). Hierzu gehört auch in den Texten belegtes indeklinables *perplex*[86] (standardsprachlich adaptiert als deklinierbares *perpleksny*).

Belegt ist mehrfach die Vorsilbe *prědk-* statt standardsprachlichem *pśed-*: *prědkgrono* statt *pśedgrono*, *prědkmě* statt *pśedmě*, *prědknjasony* statt *pśednjasony*, *prědkpołožony* statt *pśedpołožony*, *prědkstajenje* statt *pśedstajenje*, *prědkstajony* statt *pśedstajony*, *prědkstojaś* statt *pśedstojaś*, *prědkweześe* statt *pśedewześe*. Hinzu kommt die Lehnprägung *prědkbilda*, der standardsprachliches *pśikład* entspricht.

Die Verwendung von *połudnj-* statt *pódpołdnj-* in *połudnjowosłowjański* ist nach der aktuellen Wörterbuchbeschreibung ebenfalls eher umgangssprachlich. Hier scheint jedoch die Kodifikation gegen die eigentliche Verbreitung (auch im älteren gehobenen Schrifttum, bspw. in der Bibelausgabe von 1868) erfolgt zu sein.

4.2.2 Obersorbisch

4.2.2.1 Lautliche Abweichungen

Dialektal oder umgangssprachlich motivierte Abweichungen von der Standardsprache in den obersorbischen Texten sind vorwiegend durch den Sprachgebrauch im katholisch geprägten Teil der Oberlausitz bestimmt. Dem Schleifer Dialekt zuzuordnen sind demgegenüber die Formen *kólesko* (*kolesko*), *dźěćetko* (*dźěćatko*), *jarobinka* (*wjerjebinka*), *Rowniske* (*Rownjanske*), dem Gebrauch des Bautzener Dialekts entsprechende Formen wie *teho* (*toho*), *sameho* (*samoho*) sind selten. Das dürfte daran liegen, dass die Monatszeitung der evangelischen Obersorben „Pomhaj Bóh“, für deren Sprachform zum Teil spezifische Formen des Bautzener Dialekts charakteristisch sind, nicht in der vom Domowina-Verlag übermittelten Textsammlung enthalten ist. Analoges gilt hinsichtlich der für das Monitoring derzeit noch nicht zugänglichen katholischen Wochenzeitschrift „Katolski Posoł“, in der sich zum Teil sprachliche Besonderheiten des katholischen Dialekts widerspiegeln.

4.2.2.1.1 Vokale

- *'o* statt *'e* im Auslaut: Im Material finden sich die Wortformen *přebywanjo*, *přećo*, *prědowanjo*, *přewzaćo*, *wašnjo*, allerdings sind sie alle meist belegt in Os-16, einer Anthologie von Beiträgen aus historischen Publikationen, die nicht konsequent dem heutigen Standard angeglichen wurden. Der Beleg *trjenjo* in SN bezeugt jedoch die Gebräuchlichkeit dieser Formen auch in der Gegenwart.
- *ě* statt *e*: *trěnjemi* statt *trjenjemi*
- *e(j)* statt *ě*: *póžreje* statt *póžrěje*, *wumjeńk* statt *wuměnk*
- *ó* statt *o*: *jutróńčku* statt *jutrońčku*; die folgenden Belege sind wohl eher nicht als dialektal zu werten: *podróstom* statt *podrostom*, *rozróst* statt *rozrost*, *porósta* statt *porosta* – in An-

[85] Siehe hierzu den Beginn von Kap. 4.3.2. Die Aufzählung findet sich in der Internet-Datei im Abschnitt K-24.

[86] Besser geschrieben als *perpleks*.

lehnung an das einsilbige *róst*; *molemórom* statt *molemorom*, *mórowy* statt *morowy*, *brjóhi* statt *brjohi*. Die letzten drei Fälle sind auf einen Ausgleich des Wortstamms im Paradigma zurückzuführen.
- *o* statt *ó*: *nabočneje* statt *naboćneje*, *zamoženje* statt *zamóženje* (Einfluss der Aussprache von unbetontem *ó* als offenes [ɔ])
- Ausbleiben des *a-e*-Wandels zwischen palatalen Konsonanten infolge Ausgleichs des Paradigmas: *samoćahnjene* statt *samoćehnjene*, *lanjacym* statt *lehnjacym* (mit ausgelassenem stummem *h*), *wučerjaj* statt *wučerjej*, *najgrošaj* statt *najgrošej*, *guašami* statt *guašemi*
- Auf die Umgangssprache zurückzuführen sind wohl Formen wie *Luizinymi*, *Terezineho* statt *Luizynymi*, *Terezyneho* – Hintergrund ist die Verallgemeinerung des Possessivsuffixes *iny* ohne Beachtung des phonetisch harten Charakters von [z]. Analoges ist auch sonst im Obersorbischen zu beobachten bei Formen des Nom. Pl. mask. in Bezug auf Personen (*cuzi*, *lózi* statt *cuzy*, *lózy*) sowie in entsprechenden Steigerungsformen (*lóziši*, *horciši* statt *lózyši*, horcyši, vgl. Scholze 2008: 87).

4.2.2.1.2 Konsonanten
- Zusammenfall von <sć> und <šć>, Aussprache als [ʃʲʧʲ]: Damit sind Schreibungen wie *pjašću* statt *pjasću*, *košćow* statt *kosćow* zu erklären bzw. hyperkorrekte Schreibungen wie *měšćanosta* statt *měšćanosta*, *horsć* statt *horšć*, *čěšćiny* statt *čěšćiny*.
- *f* statt *hw* im Wortanlaut:[87] *fišk*, *fiškom* statt *hwižk*, *hwižkom*
- Depalatalisierung: *popjerowy* statt *popjerjowy*, *Smolera* statt *Smolerja*, *Šwicarej* statt *Šwicarjej*; einer hyperkorrekten Gegentendenz könnte man Formen zuschreiben wie *Mjezyborja* statt *Mjezybora*, *Multicarja* statt *Multicara*, *Ledźborja* statt *Ledźbora.*
- Ausbleiben der Palatalisierung in Flexion oder Wortbildung: *Sprjewinodołskeho* statt *Sprjewinodólskeho*, *skałje* statt *skale*

4.2.2.1.3 Dialektal bedingte Schnellsprechformen
- *připońca* statt *připołdnica*, *chójnčka* statt *chójnička*/*chójčka*, *twarohwe* statt *twarohowe*, *namkali* statt *namakali*

4.2.2.2 Abweichungen in der Flexion

Auch im obersorbischen Material fand sich eine Vielzahl von Belegen, die zwar größtenteils Wortgrundformen zuzuordnen wären, die schon in der Lex-DB enthalten sind, jedoch nicht den im morphologischen Generator hinterlegten (aus der bisherigen Kodifikation in Wörterbüchern und Grammatiken[88] abgeleiteten) grammatischen Flexionsregeln entsprechen. Ob es sich dabei um tatsächlich falsch gebildete Formen handelt oder ob Fehler oder Lücken in der Beschreibung bzw. Kodifizierung anzunehmen sind, muss jeweils separat beurteilt werden.

Am häufigsten sind hier „Verstöße“ gegen die aspektgesteuerte Bildung des Prozessualpartizips (nur von ip-Verben) und der Transgressive (Transgressiv des Präsens von ip-, Transgressiv des Perfekts von p-Verben). Im untersuchten Material fanden sich 36 Formen des Prozessualpartizips von einem perfektiven Verb (u. a. *dorosćaceje*, *postrowjace*, *přeběžaceje*, *přiběžacy*, *přiduce*, *přijěduce*, *připućowacy*, *přispěchacy*, *rozběžacu*, *samowukubłaca*, *skónčacy*, *skonstituowacej*, *sposrědkowace*, *wobrěčacy*, *wotjěducych*,

---

[87] Die entsprechende Aussprache ist dokumentiert in SSA 9: 218 f.

[88] Für den obersorbischen morphologischen Generator bilden hier Fasske 1981 und Faska 2012 die Basis, darüber hinaus in Fällen, die dort nicht hinreichend beschrieben sind, auch Jakubaš 1954.

wukubłace, *wužadace*, *zadźiwaca*, *zaklinčacy*, *zanjechacymi*, *započacym*, *zrudźacy*), 17 Transgressive des Präsens von perfektiven Verben (u. a. *přirunajo*, *přizjewjo*, *spóznajo*, *wučerpajo*, *zahrajo*, *započejo*, *zaručo*, *wobhladajo*) und fünf Transgressive des Perfekts von imperfektiven Verben (*lubiwši*, *poćahowawši*, *přepytowawši*, *zaběrawši*, *začuwawši*). Die hohe Frequenz derartiger der Kodifikation widersprechender Formen weist auf einen bestehenden Forschungsbedarf im Zusammenhang mit dem obersorbischen Aspekt hin, zumal manche der gefundenen Formen im DOW 1989/91 als Äquivalent angeführt werden (*přirunajo* oder *zrudźacy*) und auch im obersorbischen Textkorpus Hotko häufig nachweisbar sind (*přirunajo* 77-mal, *zrudźacy* 210-mal, *přijěducy* 161-mal, *wukubłacy* 21-mal, *přiducy* 10-mal, *přiběžacy* 10-mal).

Regelwidrig gebildete Tempusformen im Material weisen dagegen auf Unsicherheiten in der Formenbildung hin (*wobchowaše* statt *wobchowa*, *wohladaše* statt *wohlada*, *zmištrowaše* statt *zmištrowa*, *njenućichu* statt *njenućachu*, *pčołarjeja* statt *pčołarja*, *wěsćeja* statt *wěšća*, *zakónčeli* statt *zakónčili*).

Einen dialektalen Hintergrund haben Präsensformen in der 3. Pers. Pl. von Verben der *e*-Konjugation mit der Endung *-eja*: Im ausgewerteten Material fanden sich die Formen *dźeja* statt *du*, *hrimoceja* statt *hrimocu*, *tikoceja* statt *tikocu.* Allerdings stehen *hrimoceja* und *tikoceja* für ein weithin verbreitetes Bildungsmuster, während die kodifizierten Formen auf *-u* eher als veraltet anzusehen sind. Umgangssprachlich sind auch die Formen *předaja*, *wudaja*, *zapodaja* (3. Pers. Pl. Präs.) statt *předadźa*, *wudadźa*, *zapodadźa* mit erhaltenem konsonantischem Stammauslaut.

Abweichungen im Gen./Dat. Sg. mask./neutr. der Adjektive und Pronomina (*teho*, *tuteho*, *sameho*, *temu*, *samemu*, *ničomu* statt *toho*, *tutoho*, *samoho*, *ničemu*) wurden in 30 Fällen registriert, meist handelt es sich aber um Belege aus Zitaten aus älteren Texten. Die genannten Pronominalformen auf *-eho*/*-emu* entsprechen hier der schriftsprachlichen Norm des evangelischen os. Schrifttums (vor 1945), die auf dem Bautzener Dialekt fußt. Sie werden auch heute noch in religiösen Schriften der evangelischen Obersorben verwendet, so in der Zeitschrift „Pomhaj Bóh“, die nicht zum ausgewerteten Textkorpus gehört (s. 4.2.2.1). Zwei Belege für *tuteho*, *tutemu* aus „Serbske Nowiny“ dürften aber eher als Normverstoß infolge Angleichung an die Formen der Adjektivdeklination zu werten sein.

Die häufigste Abweichung vom Formenbestand der Lex-DB im Bereich der Substantivdeklination findet sich im Nom. Pl. der Maskulina. Es wurden folgende nonpersonale Formen registriert für Substantive, die in der Lex-DB nur als Rationalia registriert sind: *rowdyje*, *profije*, *cowboyje*, *muže*, *klasikarje*, *pólcaje*, *kumple*, *partnery*, *knjezy*, *demony*, *hobry*, *swědki*, *pjeraški*, *rumpodichi* – zum Teil sind solche Formen durch den Gebrauch in der Umgangssprache motiviert (*stare knjezy* statt *stari knjezojo*), zum Teil liegen jedoch zusätzliche Bedeutungen vor, wo nonpersonale Formen auch standardsprachlich normgerecht sind, wie z. B. in *sněhowe muže*; *skały, kaž hobry do krajiny stajene*; *zo so archeologiske swědki njezhubja*; *filmowe klasikarje*. Hier sind Korrekturen in der Lex-DB notwendig. Auch für *rowdy*, *profi*, *cowboy*, *pólcaj*, *kumpl*, *demon*, *pjerašk*, *rumpodich* ist wohl noch zu untersuchen, ob nonpersonale Pluralformen angesichts ihrer Frequenz im os. Textkorpus Hotko tatsächlich dem Nonstandard zuzurechnen sind.

Ähnlich wie im Niedersorbischen finden sich auch im obersorbischen Material für Substantive mit phonologisch weichem Stammauslaut normwidrige Genitivformen auf *-y* aus der Deklination der hartstämmigen Substantive: *Helly*, *Hily*, *Kruschy* statt *Helle*, *Hile*, *Krusche.* Als falsch gebildet abgelehnt werden die Flexionsformen *Lipoju* statt *Li-*

poji (mask. statt fem.), *Huskeje* statt *Huski* (adjektivische statt substantivischer Deklination), *Prožyma* statt *Prožymja* (Depalatalisierung), die Form *Połčnic* statt *Połčnicy* könnte umgangssprachlichen Hintergrund haben – die Flexion von Ortsnamen auf *-ica* wird häufig mit derjenigen von pluralischen Patronymika auf *-icy* verwechselt. Normwidrig sind auch die ohne Stammerweiterung gebildeten Wortformen Gen. Sg. *symja* statt *symjenja* und Dat. Sg. *skoću* statt *skoćeću*.

Bisher nicht registrierte vom vollen Wortstamm gebildete Flexionsformen wurden bei einigen Fremdwörtern auf *-er* verzeichnet: *paterom*, *psychiaterojo*, *bunkeru*, für die bisher nur die Deklination mit verkürztem Wortstamm angegeben wird (*patrom*, *psychiatrojo*, *bunkru*). Es könnte sich allerdings auch um ad hoc gebildete, am Nominativ orientierte Formen handeln.

Dialektalen oder umgangssprachlichen Hintergrund haben Lokativformen wie *dobroći*, *jałmožni*, *Lanzy*[89] statt *dobroće*, *jałmožnje*, *Lanze*; ebenso Dativformen für Neutra, die aus der Deklination der hartstämmigen Maskulina übertragen sind wie *pismej*, *elektrokolesej* statt *pismu*, *elektrokolesu.* Dasselbe gilt für die Numeralform *dwajo*, die statt der korrekten Form *dwaj* in Bezug auf männliche Personen verwendet wird.

Ergänzungen zum Formenbestand schon in der Lex-DB enthaltener Lexeme erfordern die Belege *hunčeća* (Gen. Sg. zu *hunčo*, bisher entsprechend PS 2014 nur ohne Stammerweiterung in der Lex-DB registriert, vgl. aber DOW 1989/91 mit Stammerweiterung), die endungslosen Gen.-Pl.-Formen *blid*, *třećin* zu *blido*, třećina (Gen. Pl. auch regulär *blidow*, *třećinow*),[90] zusätzliche neutrale Adjektivformen auf *-o* (regulär bei diesen Adjektiven auf *-'e*) (*tužno* – als Prädikatsnomen, *žiwo* – in der Redewendung *za žiwo*), zahlreiche Adverbial- und Steigerungsformen von Adjektiven, für die dies bisher in der Lex-DB nicht vorgesehen ist (*dušepastyrsce*, *fachowje*, *proeuropsce*, *inkluziwniši*, *najeficientnišo*, *najprestižniše*), vgl. 4.3.1.3. Steigerungsformen von eigentlichen Partizipien weisen wiederum auf ihre Lexikalisierung als Adjektive hin (*koncentrowanišo*, *najnjeznaćiša*, *najpožadaniša*, *njekomplikowanišo*, s. auch Fußnote 106). Die Lokativformen *Fabianu* und *zahonu* (statt *Fabianje*, *zahonje*) belegen indes, dass die Lokativendung *-u* bei den Maskulina weiter verbreitet ist, als in den Grammatiken angegeben (Einschränkung auf Substantive mit *-r* im Auslaut, so z. B. Faska 2012: 151).

4.3 Neue Lexik

4.3.1 Obersorbisch

Hier erfolgt eine erste Analyse der ermittelten bisher nicht in der Lex-DB registrierten appellativischen Lexeme. Unberücksichtigt bleiben Orts- und Personennamen sowie

---

89 Derartige auf dem katholischen Dialekt und der Umgangssprache basierende Formen (vgl. Faska 1998: 204; Scholze 2008: 369 ff.) sind in der dem Monitoring derzeit noch nicht zugänglichen Zeitschrift „Katolski Posoł“ weitaus häufiger zu beobachten.

90 Formen des Gen. Pl. auf *-i* von Maskulina mit weichem Stammauslaut, wie im Niedersorbischen bezeugt und auch für das Obersorbische durchaus wahrscheinlich, finden sich unter den nicht automatisierten Tokens im os. Material keine. Das ist auch nicht möglich, da diese Formen im os. morphologischen Generator berücksichtigt sind – entsprechende Belege wurden daher automatisch lemmatisiert.

(anders als im Niedersorbischen) deren Ableitungen (Bewohnernamen wie *Aalenčan*, *Kölnjan*, *Hochožan*, von Ortsnamen abgeleitete Adjektive wie *Bambergski*, *Černobylski*, *Drježdźansko-Mišnjanski*, Possessive zu Personennamen wie *Albertowy*, *Hancyny*, *Bałcarjowy*). Die entsprechenden Lexeme sollen, soweit sie regelgerecht gebildet und orthografisch korrekt sind, zunächst in die obersorbische Lex-DB integriert werden und später für den Aufbau der Namen-Datenbank für das Obersorbische verwendet werden. Von den 5783 verbleibenden Lexemen sind 1463 Adjektive, 379 Verben, 3096 Substantive, 74 andere (Adverbien, Partikeln, Pronomina, Präpositionen, Interjektionen), 456 Abkürzungen. Die übrigen sind nichtsorbische Wörter innerhalb von Namen von Institutionen, Medien, Musikgruppen u. Ä. (*Awful Noise*, *Crying Blue*, *Yawuru Community*) oder Teile fremdsprachlicher Wendungen (*dies academicus*, *advocatus diaboli*). Nicht alle ermittelten Lexeme sind hinsichtlich ihrer Wortbildung gelungen (*železnowobručny* statt *železnowobručaty*, *nowotwarić* statt *nowy*/*nowu*/*nowe twarić*, *wotsfalšowany* statt *wotfalšowany*, *kulturnowědnik* statt *kulturowědnik*), mitunter werden – wohl aus Unkenntnis schon etablierter und integrierter Lexeme – neue, zum Teil hinsichtlich der Wortbildung misslungene Varianten gebildet (*wjacehódnotny dawk* statt *nadhódnotowy dawk* ‚Mehrwertsteuer', *serbskoréčacy* statt *serbskoréčny* oder *serbsce réčacy*, *zakładošulski* statt *zakładnošulski*, *onomatopoetija* statt *onomatopeija*, *rohizna* fem. ‚Horn [Material]' statt *rohowina*, *Sewjerorynska-Westfalska* statt *Sewjerorynsko-Westfalska*).

Neben Wörtern des allgemeinen Sprachgebrauchs finden sich mitunter auch ungewöhnliche und kreative Wortschöpfungen bzw. ungewohnte Binnengroßschreibungen. Das gilt z. B. für das Adjektiv *appsolutny* (aus einem Musiktitel), eine Abwandlung des Adjektivs *absolutny* in Anlehnung an das Kurzwort *app* für *application*, oder für das Adjektiv *klasilektriski*, kombiniert aus *klasiski* und *elektriski*, sowie für *PowerSerb*, *QuizSerb*.

4.3.1.1 Verben

In den analysierten obersorbischen Texten wurden bisher nicht in der Lex-DB verzeichnete Wortformen von 379 unterschiedlichen Verben gefunden.[91] Von diesen Verben sind 161 imperfektiv, 212 perfektiv und bei sechs liegt wahrscheinlich Biaspektualität vor: die Lehnübersetzungen *podmolować*, *předramatizować*, *předychać*, *přiabonować*, *reintegrować* sowie *wotnajeć*, für das eine nichtnormative Flexionsform registriert wurde. Verben auf *-ować* (*e*-Konjugation) sind die häufigste Verbklasse (160 Verben). Erwartungsgemäß ist hier der Anteil von Internationalismen sehr hoch, er macht gut die Hälfte dieser Verben aus (z. B. *agitěrować*, *archiwować*, *awizěrować*, *bouncować*, *dekonstruować*, *dekontaminować*, *twitterować*). Mitunter kommen Varianten des Typs *fragmentěrować*, *fragmentować* ‚fragmentieren' vor. Häufig vertretene Verbklassen sind außerdem die *a*-Konjugation (89-mal, z. B. *domotać*, *dowudobywać*, *dowuwiwać*, *krakotać*, *kyrkać*, *potamać*, *předychać*, *spósłać*, *wobčitać*, *wobpřijimać*, *wujakotać*, *doprajeć*, *dozhotowjeć*) und *i*-Konjugation mit Infinitiv auf *-ić* (76-mal, z. B. *dokuzłarić*, *dowjertolić*, *dowujasnić*, *napakosćić*, *napomnić*, *narjejić*, *pastyrić*, *počušlić*). In der zuletzt genannten Gruppe

[91] Bemerkenswert ist, dass bei den Verben der Anteil derjenigen, die im DOW 1989/91 verzeichnet sind, im Vergleich zu den anderen Wortarten mit fast 40 % überdurchschnittlich hoch ist.

sind von Berufsbezeichnungen abgeleitete Verben ein auffällig häufig vertretenes und lexikografisch bisher nur lückenhaft erfasstes Modell (*dokuzłarić*, *fararić*, *hosćencarić*, *hribarić*, *pastyrić*, *pohončić*).

49 registrierte nicht automatisch lemmatisierbare Verbformen entsprechen nicht der kodifizierten grammatischen Norm, dabei handelt es sich zum großen Teil um die in Kapitel 4.2.2.2 angesprochenen „Verstöße" gegen die aspektgesteuerte Bildung des Prozessualpartizips (nur von ip-Verben) und der Transgressive (Transgressiv des Präsens von ip-, Transgressiv des Perfekts von p-Verben). Seltener gibt es nicht normgerecht gebildete finite Verbformen, die zum Teil umgangssprachlichen Hintergrund haben (*dźeja*, *hrimoceja*,[92] *předaja*, *tikoceja*, *wudaja*, *zapodaja*) und zum Teil wohl als Fehler zu werten sind (*njenućichu*, *přepytowa*, *wobchowaše*, *wohladaše*, *wotraže*, *woznamjenjetej*, *zmištrowaše*).

In anderen Fällen erweisen sich als neu registrierte Verbformen als bisher im morphologischen Generator nicht berücksichtigte Flexionsbesonderheiten. So zeigt das orthografisch fehlerhafte *njepočeće* statt *njepoćeće*, dass das Verb *ćec* ‚fließen' wie andere Bewegungsverben synthetische Futurformen mit *po-* bildet. Weitere solche Fälle sind *njechcyj* statt *nochcyj* in der Redewendung *chcyj njechcyj* ‚nolens volens' und die veralteten Formen *wusuže* (3. Pers. Sg. Prät. zu *wusunyć*), *zawru* (3. Pers. Pl. Präs. zu *zawrěć*), *zepěrajcy* (Bildungsvariante des Transgressivs der Gegenwart zu *zepěrać*). Die Partizipialform *wuležacymaj* ist dagegen ein Beleg dafür, dass es zum perfektiven transitiven Verb *wuležeć* ‚durchliegen, z. B. ein Bett' ein imperfektives intransitives Homonym gibt, das als Lehnübersetzung des deutschen Verbs *ausliegen* ‚hinterlegt, zugänglich sein, von Schriftstücken' entstanden und im DOW 1989/91 als Neuprägung verzeichnet ist.

Bemerkenswert ist eine Reihe von Verbformen, die entweder als nichtstandardsprachliche Flexionsformen von Verben der *i*-Konjugation[93] oder als Formen seltener Suffigierungen mit *-a-*/*-e-* erklärbar sind (z. B. *hospodarjeja* zu *hospodarić* oder suffigiertem **hospodarjeć*). Solange jedoch die entsprechenden suffigierten Verben nicht auch in anderen Flexionsformen nachweisbar sind, scheidet diese zweite Erklärung aus.

---

[92] Die Formen *tikoceja*, *hrimoceja* weisen auf ein Problem der Kodifikation hin: Verben auf *-otać* können nach der *a-* und der *e*-Konjugation flektiert werden, jedoch ist die Form der 1. Person Sg. und der 3. Pers. Pl. Präs. der *e*-Konjugation veraltet. Statt der Formen *tikocu* und *hrimocu* werden in der Schriftsprache die Formen aus der *a*-Konjugation (*tikotam*, *hrimotam*, *tikotaja, hrimotaja*) bevorzugt, in der Umgangssprache werden in der 3. Pers. Pl. jedoch die registrierten Formen auf *-eja* gebraucht.

[93] Es handelt sich außer der genannten um die Formen der 3. Pers. Präs. *pčołarjeja*, *pisanjeja*, *wěsćeja* (statt *wěšćeja*), *zwrěšćeja* sowie die Transgressive *křiwdźejo*, *ničejo*, *spomnjejo*, *tworjejo*. Formen der 3. Pers. Pl. Präs. auf *-eja* in der *i*-Konjugation sind für den katholischen Dialekt nach SSA 12: 283 mit geringer Frequenz belegt. In der auf diesem Dialekt basierenden Umgangssprache der Gegenwart sind sie jedoch offenbar nur bei Verben mit *-s/z-* im Stammauslaut verbreitet (vgl. Scholze 2008: 385, 403). Entsprechende Transgressivformen registrieren weder SSA noch Scholze 2008. Ob angesichts dessen dialektaler Hintergrund für die genannten Formen in Frage kommt oder ob sich hier Unsicherheiten in der Formenbildung äußern, bleibt vorerst offen, hier wären weitere Untersuchungen erforderlich. Außerdem ist das *l*-Partizip *zakónčeli* bezeugt, das theoretisch einer Tendenz im Bautzener Dialekt entsprechen könnte (SSA 12: 190 ff.), was aber mit Blick auf die Herkunft des Autors (SN/mwe) wenig wahrscheinlich ist.

4.3.1.1.1 Präfigierung

Perfektive Verben begegnen überwiegend als Präfigierungen. Als häufigstes Präfix ist *z(e)-*/*s-* zu verzeichnen, z. B. in *schrěnić*, *scuzbnić*, *sflankować*, *zbórbotać*, *zdeklasěrować*, *združić*, *zesłowjanšćić*, *zesunyć*, *zeznajomnić* (42 Verben). Es zeigt sich, dass auch Internationalismen regelmäßig mittels Präfigierung an das obersorbische Aktionsarten- und Aspektsystem angepasst werden (*dosaněrować*, *překalkulować*, *wufiltrować*, *sfolklorizować*, *zindustrializować*). Weitere produktive Präfixe sind *na-* (19), *pře-* (19), *do-* (19), *wu-* (16), *za-* (10), und *při-* (8): *načerpać*, *přebasnić*, *dosaněrować*, *wubagrować*, *zahudźić*, *přijuskać*. Doppelte Präfigierung ist zu beobachten in Kombination mit *do-* und *z-* (*donawuknyć*, *dowobdźěłować*, *dowopisać*, *dowudobywać*, *dozakóńčić*, *dozhotowjeć*, *znaćahać*, *zwosadźeć*, *zwukopować*).

Es wurden zwei verbale Komposita festgestellt, beides Lehnübersetzungen aus dem Deutschen: *runostajić* ‚gleichstellen' und *dalokoposłužować*[94] ‚fernbedienen', wobei das zweite wohl eine Ad-hoc-Bildung ist (kein Beleg in Hotko).

4.3.1.1.2 Suffigierung

Als einziges perfektives Verb ist *křipnyć* (zu ip *křipić* ‚knirschen') mittels Suffix abgeleitet, hinzu kommen die mit dem Suffix *-ny-* morphologisch adaptierten umgangssprachlichen Lehnwörter *šafnyć* und *šluknyć* (Letzteres möglicherweise eine individuelle Bildung), alle anderen suffigierten Verben sind imperfektiv. Häufiger sind durch Suffigierung von präfigierten perfektiven Verben abgeleitete imperfektive Verben zu beobachten (z. B. *dokładować*, *doprajeć*, *doprašować*, *dowobdźěłować*, *nałožeć*), als Suffixe treten hier *-ać*/*-eć* und *-ować*, einmal auch *-wać* (*zanjechawać*) auf. Die im Monitoring registrierten suffigierten Verben *wuchowować* und *zachowować* erfordern eine eingehendere Untersuchung. Im DOW 1989/91 sind sie systematisch als imperfektive Entsprechungen zu als perfektiv markiertem *wuchować*, *zachować* bzw. *wobchować* verzeichnet, konkurrieren aber in der Schriftsprache mit Formen dieser nichtsuffigierten Verben, die auf deren (bisher nicht berücksichtigte) Biaspektualität hinweisen (vgl. Belege aus Hotko wie *zachowajo*, *wuchowace*, *wobchowaše*). Möglicherweise verhält es sich also ähnlich mit ihnen, wie mit den bei Fasske (1981: 192) genannten präfigierten Ableitungen von Verben mit dem Suffix *-ować* (z. B. ip *molować* → p/ip *wotmolować*). Bei einigen weiteren Verben wäre zu untersuchen, ob eine Suffigierung normgerecht ist, obwohl Fasske (1981: 191) sie ablehnt (z. B. *zaržować*, *wulězować*, *zalězować*).

4.3.1.2 Substantive

Von den 3096 nicht automatisch lemmatisierbaren Substantiven, die nicht in die Kategorien der Orts- und Personennamen und ihrer Ableitungen fallen, sind 1123 Eigennamen anderer Kategorien oder deren Bestandteile. Hier gibt es zum Teil typografische Besonderheiten wie Binnengroßschreibung, wie weiter oben schon vermerkt (*LandArt*, *MinoritySafePack*, *MotoGP*, *nAund-liveband*, *JugendKlubKulTour*, *SerbskiKonsum*), verfremdende Bindestrichschreibung wie in *Con-tact* oder Versalienschreibung wie

[94] Das Verbalsubstantiv *dalokoposłužowanje* mit der Bedeutung ‚Fernbedienungsgerät' ist dagegen schon im DOW 1989/91 verzeichnet.

JANKAHANKA. Bei diesen Namen bzw. Namensbestandteilen, die häufig zum Teil fremdsprachlich sind, ist die Feststellung des Genus mitunter nicht möglich (*wubědźowanje Immoregio*, *jazzowe duwo LeDazzo*, *w* […] *rěči kikuyu*). Abgesehen von den neu registrierten Pluraliatantum (z. B. *3D-nawoči*, *awtowiki*, *biowiki*, *dźesaćtysacy*, *fotopasle*, *prawizna*, *regularije*) gehört der größte Teil der Substantive, für die das Genus festgestellt wurde, zu den Maskulina (1311), etwas weniger sind Feminina (832), von den 300 Neutra sind fast die Hälfte Verbalsubstantive (137).

Ähnlich wie bei den Adjektiven finden sich auch bei den Substantiven Reihenbildungen mit Ziffern bzw. Numeralien, die für die Lex-DB von Interesse sind. Vor allem geht es dabei um Bezeichnungen von Jubiläen (*200ćiny*, *65ćiny*, *pjećašěśćdźesaćiny*), weitere Wörter mit Ziffern sind Komposita mit den Erstgliedern *2plus-* und *3D-* (*2plus-kubłanje*, *2plus-problem*, *2plus-rjadownja*, *3D-ćišć*, *3D-fashiondesign*, *3D-laser-sken*).

Bisher nicht registrierte Substantive können in seltenen Fällen auf Phraseologismen hinweisen, wie das der Fall ist bei *chlěb a trěb*, *z nuzu a huzu*, *mětk a statk*.

Andererseits finden sich auch Neubildungen, die als reine Lehnübersetzungen aus dem Deutschen an die Stelle von etablierten sorbischen Lexemen treten, wie z. B. *zelenopysk* ‚Grünschnabel' statt *wózhriwy pjersk*, *wózhriwc*, *mlokač* (vgl. DOW 1989/91), *samonakładnistwo* ‚Selbstverlag' statt *samonakład*, *samopředstajenje* ‚Selbstvorstellung' statt *sebjepředstajenje*, *zahrodnikar* ‚Gärtner' statt *zahrodnik*.

Ähnlich wie bei den Personen- und Ortsnamen, wenngleich seltener, finden sich kopulative Komposita mit Binnenflexion, zum Teil bei Namen anderer Kategorien (*Bosniska-Hercegowina*, *Sprjewja-Nysa*, *Kamjenica-Siegmar*, *Pirna-Copitz*), aber auch bei Appellativa (*traktor-trawusyčak*, *młynk-chodotnik*, *łuka-biotop*). Für diesen Typ müssen im morphologischen Generator noch Lösungen gefunden werden, um die korrekten Flexionskombinationen zu bilden.

Unter den Substantiven finden sich zahlreiche Internationalismen (*akronym*, *antropozofija*, *dendrochronologija*, *digitalizat*, *ewaluator*, *sublicenca*, *transkulturalita*) und Anglizismen. Im Unterschied zu Ersteren werden Letztere meist orthografisch nicht adaptiert (*backpacker*, *bobbycar*, *chutney*, *crowdfunding*, *emoji*), was aber nicht ausgeschlossen ist (z. B. *ekwipment*, *ewent*, *3D-laser-sken*, scherzhaft auch *hajwej* ‚Highway'). Entlehnungen aus dem Deutschen sind entsprechend der puristischen Strategie der obersorbischen Schriftsprache nicht so häufig, dennoch fehlen sie durchaus nicht. Zum Teil handelt es sich um Wörter, die schon im DOW 1989/91 als umgangssprachlich oder veraltet/veraltend registriert sind (*handler*, *hantwjerski*, *kepnjenje*, *šabernak*, *špikcedlka*, *štepsl*, *tyšer*, *tyšernja*). Bisher lexikografisch nicht registriert wurden z. B. *bogenlampa*, *buchtla*, *dejner* ‚Döner', *drona* ‚Drohne', *florentiner*, *klamawk*, *knaker*, *lojta*, *randala*, *reichsbürger*, *reichsmarka*, *sajblink*, *spint*. Außerdem gibt es auch morphologisch adaptierte und zum Teil flektierte Entlehnungen aus anderen Sprachen[95], vgl. z. B. *majówka* (poln.), *balokopać*, *šorca*, *cypjel*, *kołac*, *namša*, *stog*, *šuflitowanje* (ns.), *antipasti*, *balsamiko*, *bruschetta*, *bambinije* (ital.), *cantor*, *dubia*, *summa* (lat.), *guaš* (franz. *gouache*), *haikai* (japan.), *bendir*, *guellal* (arab.), *fujara* (slowak.). Darüber hinaus kommen entlehnte Elemente auch in Komposita mit obersorbischen Elementen und in Derivaten vor.

---

[95] Insbesondere für Entlehnungen aus nichtslawischen Sprachen ist in der Regel eine Vermittlung durch das Deutsche anzunehmen.

Als entlehnte Kurzwörter wurden im obersorbischen Material registriert: *app*, *frust*, *medi* (Mediziner), *prof*, *soko* (Sonderkommission).

Wohl als kreative und zum Teil scherzhafte Neuschöpfungen sind einige Substantive mit Eigennamencharakter einzuschätzen: *piwoněr* (Kofferwort aus *piwo* und *pioněr*), *SwójPuć* (aus *swójbne pućowanje*), *Dźěćiznak*[96] sowie das Appellativum *připońc* (Rückbildung als Bezeichnung einer männlichen Person aus umgangssprachlich *připońca* ‚Mittagsfrau, übertr. lästige Fragenstellerin' oder vom davon abgeleiteten umgangssprachlichen Verb *připońc(o)wać* ‚lästige oder viele Fragen stellen').

4.3.1.2.1 Derivation

Unter den durch Affigierung gebildeten neuen Formationen überwiegen im obersorbischen Material deutlich die Suffigierungen. Dabei wurden Fälle, bei denen die Ableitungsbasis präfigiert ist, nur als Suffigierungen gewertet, so z. B. bei *njezrałc* ← *njezrały* (nichtpräfigiertes **zrałc* ist nicht belegt), *předčitar* ← *předčitać*, *superintendentka* ← *superintendent*. Außer in reinen Derivaten kommen Affigierungen natürlich auch in Komposita vor (*AfD-politikar*, *ja-powědar*, *Wulkobritaničan*), darüber hinaus gibt es auch von Komposita abgeleitete Derivate (*lodohokejistka* ← *lodohokejist*, *šefredaktorka* ← *šefredaktor*), daher ist die Abgrenzung mitunter problematisch. Als Affigierungen, die nicht Komposita oder Ableitungen von solchen sind, wurden im vorliegenden Material 709 Substantive gewertet.

4.3.1.2.1.1 Präfigierung

Im Material finden sich indigene und entlehnte Suffixe wie *anti-*, *arcy-*, *eks-*, *wice-* und *super-*. Das häufigste Präfix ist ebenso wie im Niedersorbischen das Negationspräfix *nje-* (insgesamt 31 belegte Lexeme, z. B. *njeakceptanca*, *njeaktiwnosć*, *njechłostajomnosć*, *nječłon*, *njedobro*, *Njedomowinjan*). Es folgen *před-* (19), *mjezy-* (13) und *wice-* (10), z. B. *předdźěło*, *předfinisaža*, *předjědź*, *mjezyetapa*, *mjezyfacit*, *mjezygeneracija*, *wicedirektor*, *wicedirigent*, *wicehejtman*. Weitere belegte Präfixe mit mindestens drei Belegen sind *pod-*, *super-*, *bjez-*, *pra-*, *eks-* (*podčłowjek*, *superspěšnik*, *bjezbarjernosć*, *prapismo*, *eks-hrajer*[97]).

4.3.1.2.1.2 Suffigierung

Die nach den Verbalsubstantiven (s. o.) häufigste Kategorie von Suffigierungen sind die Feminativa, Bezeichnungen von weiblichen Personen, meist als Pendant zu entsprechenden Maskulina. Im bearbeiteten Material wurden 101 solche Bezeichnungen registriert (*bloggerka*, *oligarchka*, *premierka*, *propstowka*, *protagonistka*, *rabinerka*, *repetitorka*, *restawratorka* u. a.), zum Teil handelt es sich um Derivate von Komposita (*lodohokejistka*, *profihrajerka*, *samomordarka*, *šefredaktorka*, *sobuorganizatorka*, *sobupožadarka*, *solo-travellerka*, *TV-žurnalistka*). Die meisten sind mit dem Suffix *-ka* gebildet, sechs mit der Suffixkombination *-owka* (*antropologowka*, *etnografowka*, *fotografowka*, *lotsowka*,

---

[96] Name der Kinderbeilage der „Serbske Nowiny" (SN), angelehnt an *Předźeznak* (Silvesterbeilage von SN), was wiederum ein Kofferwort aus *Předźenak* ‚Garnhändler, Name der Wochenendbeilage von SN' und *znak* ‚verkehrt, auf dem Kopf stehend' ist.

[97] Korrekt ohne Bindestrich *ekshrajer* – mit regelwidriger Bindestrichschreibung sind außerdem *eks-sobudźěłaćer* und *ko-režiser* belegt.

propstowka, *wirtuozowka*), zwei mit dem Suffix *-ica* (*kulturnica*, *tankownica*), einmal ist das Suffix *-ča* belegt (*filmytwórča*).[98]

Als weitere sehr produktive Kategorie sind Deminutiva zu nennen, von denen 61 verzeichnet wurden (z. B. *apelsinka*, *awtko*, *bibliotečka*, *dubik*, *hłupačk*, *jandźelk*, *kwančik*, *małoměstačko*, *měrkušk*, *mjeńšinka*, *młodušk*, *portrećik*, *praprawnučk*, *procenćik*, *zwučk*), darunter auch doppelt suffigierte (z. B. *dubičk*, *filmčk*, *hołbičk*, *latarnčka*). Charakteristisch für das Obersorbische ist die Möglichkeit der Deminutivierung von Verbalsubstantiven (*přispomnjenčko*, *prědowančko*, *stonančko*, *woptančko*, *zybolenčko*) und deadjektivischen Abstrakta (*njewšědnostka*, *słabostka*). Augmentativa waren dagegen nur wenige zu registrieren (*dubisko*, *mróčisko*, *płomjenisko*, *štomisko*, *žabisko*).

Die häufigsten Suffixe (mit mehr als 20 Formationen) unter den im Monitoring neu ermittelten Bildungen sind *-ar* (72), *-ija* (48), *-stwo* (46), *-nosć* (42), *-nik* (41), *-er* (34), *-acija* (31), *-ak* (29), *-osć* (28), *-nišćo* (22), *-ina* (24). Die Suffixe *-ar* und *-er* bilden mit den Nomina Agentis als Bezeichnungen männlicher Personen eine weitere hochproduktive Wortbildungskategorie (z. B. *camprowar*, *dowožowar*, *dwanatkar*, *debjer*, *dowoler*, *předpłaćer*). Ein weiteres häufiges Bildungsmodell für Personenbezeichnungen etabliert das Suffix *-(n)ik* (*ABCnik*, *FSGnik*, *industrijnik*, *iniciatiwnik*). Wie zu sehen ist, können Suffigierungen auch von Akronymen gebildet werden. Unter den durch Suffigierung gebildeten Personenbezeichnungen finden sich einige, die aktuelle gesellschaftliche Entwicklungen im sorbischen Bereich widerspiegeln: *Stupdalnik* (Anhänger der Initiative „Stup dale“ in Dresden), *sejmar* und *sejmikar* (Vertreter bzw. Anhänger des „Serbski sejm“), *Satkular* (mit zugehörigem Feminativum *Satkularka*, Mitarbeiter bzw. Mitarbeiterin der Jugendradiosendung „Satkula“).

Als Beispiele für die weiteren genannten häufigsten Suffigierungen seien genannt: *akolutija*,[99] *desiluzija*, *docentstwo*, *dźiwadźelnistwo*, *coolnosć*, *lajskosć*, *bórčak*, *chmurjak*, *alfabetizacija*,[100] *charakterizacija*, *dowolnišćo*, *festiwalnišćo*, *asturišćina*,[101] *friulšćina*.

4.3.1.2.2 Komposita

Die Zahl der Komposita im ausgewerteten Material (731) übersteigt die der reinen Derivate kaum, jedoch sind auch davon viele mit Affixbeteiligung gebildet (295). Das zeigt deutlich, dass die Derivation im Obersorbischen trotz wachsender Bedeutung der Komposition auch in der Gegenwart weiter als produktives Wortbildungsmittel funktioniert (Pohončowa 2017: 72 f.).

---

[98] Als produktive Wortbildungskategorie werden solche Substantive in herkömmlichen Wörterbüchern häufig nicht verzeichnet, und so finden sich auch nur acht von den im Monitoring registrierten Feminativa im DOW 1989/91 (*iniciatorka*, *kameradka*, *kapitanka*, *kontoristka*, *připowědźerka*, *Sokołka*, *tankownica*, *wirtuozowka*).

[99] Dieses Suffix erscheint in Internationalismen, die im Deutschen auf *-ie* oder *-ion* auslauten.

[100] Das Suffix *-acija* entspricht deutschem *-ation* oder *-ierung* und konkurriert im Obersorbischen mit den Verbalsubstantiven, wie die beiden Belege *profesionalizacija* und *profesionalizowanje* verdeutlichen. Parallelbildungen zu adjektivischen Ableitungen auf *-osć* gibt es mit dem Suffix *-ita* (mit beiden Suffixen im Material belegt ist *hybridita* und *hybridnosć*).

[101] Die Belege mit dem Suffix *-ina* sind bis auf *jahlina* und *běrtlina* Bezeichnungen von Sprachen.

4.3.1.2.2.1 *o*-Komposita

Die Bildung von Komposita mit dem Bindevokal *-o-* tritt seit den puristischen Bestrebungen des 19. Jahrhunderts vermehrt im Obersorbischen auf. Das zeigt sich auch im untersuchten Material – es wurden 123 solche Formationen gefunden, wobei als Erstglieder vor allem Adjektive und Substantive auftreten, aber auch Pronomina wie *sam* oder *jenaki* (*jenakosplažnosć*) sowie das Numerale *jedyn* (*jednoréčnosć*, *jednomjenowosć*). Einige bilden umfangreiche Reihen, so wie *wulko-* (24), *nowo-* (14), *samo-* (16): z. B. *wulkodemonstracija*, *wulkoformat*, *wulkoimam*, *nowočas*, *nowoinstrumentacija*, *noworéčnik*, *samocensura*, *samomordarka*, *samošćipanje.* Als Beispiele mit substantivischen Erstgliedern seien hier genannt *dróhotwarc*, *klankohra*, *klimoškit*, *krajinotwar*, *lodosport*, *škleńcotwar*, *wnučkokmanosć*, *wodobulist*, *zynkoslěd*.

4.3.1.2.2.2 Komposita mit genitivischem Erstglied

Dieses am Deutschen orientierte Modell ist mit 32 Belegen deutlich seltener. Als Erstglied am häufigsten vertreten ist das Reflexivpronomen im Genitiv (z. B. *sebjedisciplina*, *sebjepomoc*, *sebjepostajenje*). Außerdem treten Numeralien (*dwuzornko*, *dźesaćitysacy*, *sedmištučka*, *šěsćihwězda*)[102] und Substantive auf (*dźěćidźěćo*, *jednanjakmanosć*, *ludźiznajer*, *próstwystajer*, *swětawotewrjenosć*, *wnučkowkmanosć*).

4.3.1.2.2.3 Komposita ohne Bindeglied

Dieser ans Deutsche angelehnte Kompositatyp, meist mit entlehntem Erstglied (prominentes Beispiel *šulknižki*), der im älteren Obersorbischen sehr verbreitet war (Jentsch 1999: 72 ff.), wurde durch die puristischen Bestrebungen des 19. Jahrhunderts weitgehend aus der Schriftsprache verdrängt (ebd.: 197 ff.), hielt sich aber in der Umgangssprache (Wornar 2001). In den letzten Jahrzehnten nimmt seine Frequenz jedoch auch in der Schriftsprache wieder zu, allerdings meist nicht mit deutschen, sondern eher mit anderssprachlichen, häufig englischen Erstgliedern sowie mit Eigennamen, Kurzwörtern und Akronymen (Wölke 2006: 45–47). Im Wörterbuch neuer Lexik (DOW-NL 2006) werden Formationen wie *joggingwoblek*, *babyfoodwohrěwak* zum Teil noch als umgangssprachlich markiert, nicht aber solche mit Kurzwörtern oder Akronymen als Erstglied (*profiklub*, *ABM-přistajeny*). In den ausgewerteten obersorbischen Texten machen solche Komposita über drei Viertel der Komposita aus (575), was von ihrer großen Vitalität zeugt. Es wird zu diskutieren sein, ob ihre Wertung als umgangssprachlich für die Schriftsprache der Gegenwart noch zu halten ist.

Dabei gibt es durchaus auch im Sorbischen verschiedene Kompositamodelle, die ohne Bindeglied (Bindevokal oder Kasusform) auskommen.

4.3.1.2.2.3.1 Mit indigenem Erstglied

Hier fallen zunächst Komposita mit Numeralien auf. Diese Bildungen sind standardsprachlich, so bei *połbur*, *połfinale*, *Połitalčan*, *połmarathon*, *połpolo*, *połserija*, *stoléće*, *štyričarowosć*, *třibój* (als Variante zum *o*-Kompositum *trojobój*). Ähnliches gilt für Komposita mit *wjele* (*wjelekulturnosć*, *wjeleworštowosć*), die allerdings auch als Suffigierung der adjektivischen Komposita (*wjelekulturny* bzw. *wjeleworštowy*) interpretiert werden können. Besonders produktive Erstglieder wie z. B. *sobu-* könnte

[102] Die genannten Genitivformen der Numeralien sind veraltet.

man möglicherweise als Präfixoide werten, womit solche Formationen an der Grenze zur Derivation zu verorten wären (im Material fanden sich 30 Belege mit *sobu-*; z. B. *sobuhudźbnik*, *sobukrajan*, *sobupostajenje*). Die Behandlung unter den Komposita folgt der Einordnung bei Pohončowa 2017: 74.

Als Lehnübersetzungen deutscher Vorlagen dürften folgende Belege einzuschätzen sein: *hišće-prezident* (Noch-Präsident), *ludźozwuk* (Menschenlaut), *dźesaćtysacy* (Zehntausende). Der literaturwissenschaftliche Terminus *ja-powědar* (Ich-Erzähler) sowie die Ausdrücke *kołočara* (Rundstrecke), *kołokurs* (Rundkurs) und *kołopuć* (Rundweg) sind mittlerweile wohl im schriftsprachlichen Gebrauch etabliert.

4.3.1.2.2.3.2 Mit Akronymen und Kurzwörtern

Eine große Gruppe unter den bindegliedlosen Komposita bilden solche mit einem Akronym als Erstglied. Im Material wurden 78 solche Formationen gefunden, wobei die Akronyme zum Teil den Charakter von Eigennamen haben (*FIFA-turněr*, *GiG-festiwal*, *DFB-pokal*, *IBA-terasa*, *PEN-centrum*), andere sind Abkürzungen anderer Art (*e-koleso*, *HQL-lampa*, *LED-wobswětlenje*, *VR-nawoči*). Außerdem treten Einzelbuchstaben mit Symbolcharakter (Teile von Klassifikationen) als Erstglied auf (*A-młodźina*, *B-dur*, *C-rjadownja*, *h-moll*, *Y-twarjenje*).

Kurzwörter treten in 48 Fällen als Erstglied auf, besonders häufig sind *bio-* (*biodiwersita*, *bio-kruška*, *biomjaso*), *eko-* (*ekobilanca*, *ekofarma*, *ekohysterija*) und *krimi-* (*krimiawtor*, *krimi-kniha*, *krimispisowaćel*). Weiter kommen mehrfach vor *abi-* (*abibal*, *abi-lětnik*, *abi-party*), *euro-* (*Europark*, *europarlament*, *eurozapósłanc*), *info-* (*infomobil*, *infoportal*, *inforadijo*), *profi-* (*profihrajerka*, *profikolesowar*, *profikopar*). Einzelbelege sind *šokocar*, *soli-připłatk*, *app-wobchod*, *homomandźelstwo*, *izomata*, *kombi(-)tiket*, *sat-připrawa*. Ebenfalls als Kurzwort wurde das Erstglied *2plus-* (Kurzname einer Sprachunterrichtskonzeption) gewertet (*2plus-kubłanje*, *2plus-problem*, *2plus-rjadownja*, *2plus-skupina*, *2plus-šuler*).

4.3.1.2.2.3.3 Mit fremdem Erstglied (hybride Komposita)

Besonders häufige Erstglieder sind hier: *elektro-*, *awdijo-*, *widejo-*, *online-*, *šef-* (z. B. *awdijodeskripcija*, *elektrocigareta*, *widejopinca*, *online-spěwnik*, *šefmechanikar*). Diese Reihenbildungen sind in der Standardsprache akzeptiert, sie konkurrieren zum Teil mit Syntagmen mit entsprechenden adjektivischen Derivaten (*onlinowy*, *awdijowy*, *widejowy*), die schon in der Lex-DB verzeichnet sind.

Auffällig ist auch eine Reihe von zehn Komposita mit dem Erstglied *awto-* (z. B. *awtobranša*, *awtoindustrija*, *awtomechanikarka*). Im DOW-NL sind die Komposita mit diesem Erstglied oft als umgangssprachlich markiert (z. B. *awtokino*, *awtobomba*, *awtokran*), nicht aber *awtodróha*, *awtostop*, *awtołamak*. Die gefundenen Belege erweitern diese Reihe und sprechen dafür, diese Bildungen als standardsprachlich anzuerkennen. Analoges könnte für die Komposita mit dem Erstglied *foto-* gelten (*fotomodel*, *fotopasle*, *fotowustajeńca*).

Die meisten fremden Erstglieder sind Anglizismen (z. B. *beachbara*, *blackout-wokomik*, *dumpersportowc*, *funkskupina*, *hightech-aparaćik*). Doch gibt es auch solche aus dem Lateinischen (*a-cappella-hudźba*, *solo-travellerka*), Italienischen (*parmezan-twarožk*, *brutto-zasłužba*), Spanischen (*salmorejo-juška*), Französischen (*a-la-cart-hósć*,[103] *floret-mustwo*),

[103] Korrekt wäre *à-la-carte-hósć*.

Japanischen (*no-pišćele*, *koi-ryba*), Arabischen (*shisha-bara*) und nicht zuletzt aus dem Deutschen (*hamplmuž*, *hajldomizna*, *kita-olympiada*, *Wiesenbeats-swjedźeń*), wobei auch die meisten nichtdeutschen Erstglieder wohl über das Deutsche ins Obersorbische übernommen wurden. Mitunter finden sich auch komplexe Erstglieder wie in *drive-in-hosćenc*, *drum'n'bass-hudźba*, *heavy-metal-hudźba*, *Juniortriatloncup*, *to-do-lisćina*, *toptenzaměstnjenje*.

Nicht eindeutig zu entscheiden ist in manchen Fällen, ob ein Kompositum als Ganzes als Entlehnung zu betrachten wäre, z. B. in Fällen wie *destroyed-look*, *cyber-mobbing*, *disco-fox*, *down-syndrom*, *dumperteam*, *elektrocart*.

4.3.1.2.2.3.4 Mit Eigennamen als Erstglied

Ein Teil der gefundenen Komposita hat als Erstglied einen Eigennamen (Personen-, Ortsnamen oder auch Namen von Gruppen, Institutionen oder Einrichtungen), der auch aus mehreren Komponenten bestehen kann. In der obersorbischen Schriftsprache werden solche Bildungen normalerweise vermieden, stattdessen werden Syntagmen mit Adjektiv, Possessiv, Genitivattribut oder Apposition bevorzugt. D. h. zu erwarten wäre statt *Kulow-cup – Kulowski cup*, statt *Fröbel-pěstowarnja – Fröbelowa pěstowarnja*, statt *bennett-kenguru – Bennettowy kenguru*, statt *Malte-Rogacki-band – band Malte Rogackeho*, statt *Satkula-bičtura – bičtura Satkule*, statt *Melodia-diskoteka – diskoteka Melodia.* Im ausgewerteten Material wurden 55 solche Komposita festgestellt. Man kann sie wohl nicht alle pauschal als umgangssprachlich werten. Hier werden individuelle Entscheidungen getroffen werden müssen, wie z. B. in solchen Fällen wie *legokamušk*, *Leader-region*, *brexit-party*, *Pixi-knižka*, *Trilex-ćah*. Fest in der Schriftsprache der Gegenwart verankert sind mittlerweile *Jolka-swjedźeń* und die Komposita mit *WITAJ-* (im Material durch formale Vereinheitlichung stets als *Witaj-*: *Witaj-camp*, *Witaj-dźěćo*, *Witaj-hibanje*, *Witaj-koncept*, *Witaj-kubłanje*, *Witaj-pěstowarnja*, *Witaj-rjadownja*, *Witaj-šulerka*, *Witaj-wučba*). Als zumindest salopp, wenn nicht umgangssprachlich wären wohl Fälle einzuordnen wie *Raiffeisenbanka*, *Suffolk-wowca*, *Pasternak-młyn*, *Juras-młyn*.

4.3.1.2.2.4 Zusammenrückungen

Eine besondere Gruppe an der Grenze zwischen Komposition und Derivation bilden Zusammenrückungen, die auf Syntagmen mit verbalem Kern zurückgehen, kombiniert mit Adverb, Substantiv oder Präpositionalgruppe, in der Regel als Nomina Actionis (*jejkadebjenje*) oder Nomina Agentis (*filmytwórc*), aber auch Nomina Instrumenti sind möglich (*trawusyčak*). Formationen dieses Typs, die im Deutschen eine produktive Kategorie bilden, sind auch im Obersorbischen häufig vertreten, im Unterschied zu anderen slawischen Sprachen. Das Bildungsmodell wird durchaus als schriftsprachlich angesehen (u. a. Pohončowa 2017: 73 f., Michałk 1974: 509). Im ausgewerteten Material fanden sich zahlreiche solche Bildungen mit Adverbien (*domabyće*, *jowpřińdźenje*, *napřemojězdźenje*, *nutřćehnjenje*, *tam-a-sem-jězdźenje*, *znowawobsadźenje*), mit Substantiv im Akkusativ (*filmytwórc*, *husykupc*, *jejkawóskowanje*, *kachlestajenje*, *pućetwar*, *raketymjetak*), mit Präpositionalgruppe (*do-škita-branje, na-so-čakanje*) oder mit Substantiv und Adverb (*mejudelebraće* ← *meju dele brać*). Als Zusammenrückungen sind auch *babkahłód* ‚Hungerblümchen' sowie *wšowěm* ‚Alleswisser' zu erklären.

4.3.1.3 Adjektive

Von den 1463 Adjektivlexemen sind 289 (19,7 %) schon im DOW 1989/91 registriert. Eine relativ große Gruppe (insgesamt 180) bilden hier Formationen aus Zahlen bzw. Numeralien und Adjektiven: Komposita mit Ziffern, überwiegend ohne Bindestrich geschrieben (*1000lětny*, *100dnjowski*, *12hodźinski*, *146stronski*, *15kilometrowski*, *2-hłósny*), seltener waren Ordinalia festzustellen, die als Kombination aus Ziffernfolge und Endung geschrieben sind (*1760ty*, *1930ty*, *1950-ty*). Hinzu kommen Komposita aus ausgeschriebenem Numeral und adjektivischem Zweitglied (*třiapołhektarski*, *třinaćelětny*, *štyrikrajowy*, *pjećčłonski*, *pjećsadźbowy*). Damit machen derartige Reihenbildungen ca. 12 % aller Adjektiv-Formationen aus. Die Aufnahme aller derartiger Lexeme in ein Wörterbuch im herkömmlichen Sinne dürfte problematisch sein, für die Lex-DB, die dem Monitoring zugrunde gelegt wird, aber auch für die automatische Rechtschreibkontrolle sind sie jedoch von Interesse.

Interessant ist auch ein Blick auf die Ableitungsbasen bzw. die Erst- und Zweitglieder der Komposita hinsichtlich ihrer Herkunft. Nach einer groben Klassifizierung wurde ein Anteil von rund einem Drittel festgestellt, bei dem Lehn- und Fremdwörter zugrunde lagen, ein deutlicher Hinweis auf die Intellektualisierung der Sprache insbesondere in der Publizistik, die ja den Löwenanteil der Texte ausmacht. Hier einige Beispiele: *akcionarski*, *akribiski*, *anaboliski*, *backpackerski*, *bagatelny*, *chórosinfoniski*, *crossmedialny*, *crowdfundingowy*, *etnolinguistiski*.

Zum Teil wurden Adverbialformen von Adjektiven gefunden, die schon in der Lex-DB registriert sind, jedoch ohne Möglichkeit der Bildung dieser Formen (z. B. *dušepastyrski*, *ekologiski*, *faktowy*, *gentechniski*, *sezonowy*). Hier offenbart sich ein Unterschied zwischen dem ober- und dem niedersorbischen morphologischen Generator: Während dieser die Adverbialformen als separate Lexeme wertet, sind sie bei jenem in das Adjektivparadigma integriert. Die obersorbische Lösung ist insofern problematisch, als in Wörterbüchern nicht für jedes einzelne Adjektiv angegeben ist, ob Adverbialformen gebildet werden (dasselbe gilt für die Bildbarkeit von Steigerungsformen). Die Beschreibung in den Grammatiken stützt sich auf semantisch motivierte Kategorien, für die prototypische Beispiele angegeben werden, was als Entscheidungsgrundlage nicht in jedem Fall ausreicht.[104] Die bisher nicht registrierten Adjektive *dwukolijowy*, *holistiski*, *imersiwny*, *połnozwučny*, *stoprocentowski*, *zwukotechniski* wurden sowohl in kongruenten Adjektivformen als auch in Adverbialformen registriert, nur in der Adverbialform wurden nachgewiesen: *aransce*, *nadpřerěznje*, *njetajomnje*, *politisko-ekonomisce*, *prapremjernje*, *staroruscе*, *uzualnje*, *wysokoněmsce*, *zadwělnje* (die entsprechenden Adjektive wären *aranski*, *nadpřerězny*, *njetajomny*, *politisko-ekonomiski*, *prapremjerny*, *staroruski*, *uzualny*, *wysokoněmski*, *zadwělny*). Bisher nicht in der Lex-DB enthaltene deminutivische Adverbialformen sind *ćišinko* zu *ćiše* und *rjenko* zu *rjenje*, wovon das erste sehr gebräuchlich ist (47 Belege in Hotko, auch in den historischen Wörterbüchern von Kral 1927, Pfuhl 1866 und Rězak 1920), das zweite eher selten. Als echte Adverbien neu registriert wurden das umgangssprachliche *našlak* ‚sofort, auf Anhieb', *praněhdy* ‚zu Ur-

[104] Die Fähigkeit, Adverbialformen zu bilden, ist nach Fasske 1981: 355/369 auf Adjektive beschränkt, „deren lexikalische Bedeutung einen merkmalcharakterisierenden Gebrauch zulässt", die Bildung von Steigerungsformen auf „relativ-qualitative Adjektive".

zeiten‘, *předloni* ‚vor zwei Jahren‘ sowie *častohdy*, das laut PS 2005 bzw. 2014 getrennt geschrieben werden sollte.

4.3.1.3.1 Derivate

4.3.1.3.1.1 Präfigierung

Das häufigste Präfix bei den Adjektiven ist erwartungsgemäß *nje-* (74 Lexeme, z. B. *njeaktiwny*, *njedarniwy*, *njedemokratiski*, *njedostojny*, *njehóstliwy*). Dabei wurden vier Adjektive gefunden, die bisher nur mit dem fremden Präfix *in-* registriert waren (*njeaktiwny*, *njedirektny*, *njediskutabelny*, *njekonsekwentny*), was auf eine fortschreitende Integration der entsprechenden nicht negierten Adjektive hinweist. Andere Präfixe waren eher sporadisch zu verzeichnen: *bjez-* (*bjezbojowy*, *bjezčasowy*, *bjezhódnotny*, *bjezkonfliktny*, *bjezpředmjetowy*), *pra-* (*prazastarski*, *pra-Chróšćanski*[105]). Auch fremde Präfixe sind hier produktiv: *anti-* (*anti-atomowy*, *antidopingowy*, *antimigraciski*, *antimoderny*, *antiserbski*), *pro-* (*proeuropski*, *proruski*), *ultra-* (*ultrakatolski*, *ultraprawicarski*, *ultraradikalny*), *inter-* (*interministerielny*, *internabožinski*). Belege wie *transseksualny, interaktiwny* sind wohl nicht als präfigiert, sondern eher als im Ganzen entlehnt zu betrachten, relevant ist jedoch, dass sie im Monitoring als neue Lexeme registriert wurden.

4.3.1.3.1.2 Suffigierung

Generell überwiegen bei den registrierten neuen Adjektiven komplexere Strukturen, die selbst oder deren Zweitglied (sofern es Komposita sind) mit einem Suffix abgeleitet sind. Am häufigsten sind hier folgende Suffixe (Belegzahl und Beispiele in Klammern): *-ski* (534: *anaboliski*, *antimigraciski*, *butrowohórski*, *chemisko-syntetiski*, *debatowanski*); *-ny* (413: *awtomobilny*, *bagatelny*, *cyłotowaršnostny*, *dalnowobchadny*, *e-mailny*, auch zahlreiche Passivpartizipien[106] wie *samobasnjeny*, *samodebjeny*, *samopasleny*) und *-owy* (275: *akrobatikowy*, *aronijowy*, *bjezbojowy*, *cateringowy*, *chórowy*); *-omny* (45: *dopokazujomny*, *knihujomny*, *njesłyšomny*, *njewužiwajomny*, *polěpšomny*); *-aty* (41: *běłonóžkaty*, *błyšćaty*, *dwuhłowaty*, *horbikaty*, *klimpotaty*); *-acy* (24, meist Partizipien: *dorosćacy*, *historisko-přirunowacy*, *jenakmyslacy*, *krejzaprajacy*, *napinacy*, aber auch desubstantivische Ableitungen wie *jejkacy*, *kokulacy*, *kubołčacy*, *kwětkacy*, *maćerno-dźěćacy*); *-(l)iwy* (15: *powabliwy*, *přiwabliwy*, *sobučućiwy*, *wudźeržliwy*, *wuprajiwy*).

Bezüglich der selteneren Bildungsmuster ist hinzuweisen auf drei bisher nicht registrierte Adjektive, die auf die synchron nicht mehr produktiven *l*-Partizipien zurückgehen: *wušły*, *dóšły* und *wotrostły*.[107]

---

[105] Mit ungewöhnlicher Bindestrich- und Binnengroßschreibung – vermutlich sollte umgangen werden, eine Ableitung vom Ortsnamen mit Kleinbuchstaben zu schreiben. Die Orthografieregeln des Obersorbischen sehen einen solchen Fall nicht vor.

[106] Partizipien werden im obersorbischen morphologischen Generator als Teil der entsprechenden Verbalparadigmen generiert, daher wurden sie, falls sich ein solches Verb ermitteln ließ, beim Monitoring nicht als selbstständige Adjektive behandelt. Das gilt natürlich nicht in Fällen, wo kein entsprechendes Verb zu ermitteln ist wie z. B. bei *o*-Komposita (*nowowuhotowany*, *samobasnjeny*, *módroćišćany*, *złotowušiwany*). Steigerungs- und Exzessivformen wie *diferencowaniši*, *koncentrowanišo*, *přenapinace*, *njekomplikowanišo*, *najnjeznaćiša*, *přepowučowace* zeigen, dass die ursprünglichen Partizipien als Adjektive lexikalisiert wurden: *diferencowany*, *koncentrowany*, *napinacy*, *njekomplikowany*, *njeznaty*, *powučowacy.*

[107] Vgl. dazu in der bisherigen Lex-DB *přešły*, *zašły*, *dorostły*, wobei Letzteres im DOW 1989/91 als veraltend gewertet wird.

4.3.1.3.2 Komposita

Abgesehen von den Bildungen mit Zahlen wurden 437 adjektivische Komposita registriert, vielfach (97) werden sie mit Bindestrich geschrieben, eher selten in Varianten mit und ohne Bindestrich (*duchowno(-)kulturny*, *literarno(-)wědomostny*, *měšćansko(-)twarski*, *towaršnostno(-)politiski*, *saksko(-)swěrny*). Bindestrichschreibungen überwiegen regelgerecht, wenn es um gleichranginge Farbkombinationen und Kombinationen von Völker-, Sprach- und Regionenbezeichnungen geht (*rusko-ameriski*, *saksko-čěski*, *serbisko-němski*, *serbsko-čěski*, *zeleno-běły*, *žołto-čerwjeny*).[108] Bei Kombinationen von Bezeichnungen politischer, religiöser bzw. gesellschaftlicher Strömungen lässt sich keine solche Präferenz ausmachen (vgl. z. B. *nabožnonarodny*/*narodno-nabožny*, *měšćansko-twarski*/*měšćanskotwarski*, *kulturno-politiski*/*kulturnopolitiski*[109]) – möglicherweise weist dies auf Defizite der Kodifizierung (aktuell gültiges Regelwerk in PS 1981: 632 f.) hin.

In den meisten Fällen liegen *o*-Komposita vor, die häufigsten Erstglieder sind *samo-* (18), *nowo-* (18), *staro-* (9) sowie die Bezeichnungen der Himmelsrichtungen (insgesamt 35, z. B *zapadoafriski*, *južnomorawski*, *sewjernosewjerozapadny*).[110]

Als weitere Kompositatypen sind zu verzeichnen: solche mit indeklinablen Erstgliedern wie *wjele*, *wjace*, *zwonka*, *nimo*, *sobu*, *poł* (*wjelełopjenaty*, *wjaceréčny*, *zwonkamandźelski*, *nimosłužbny*, *sobupostajowanski*, *połprofesionalny*), mit akkusativischen Erstgliedern (*hłowuwjerćaty*, *kosćełamakowy*, *lětadławotwobaranski*, *róštytwarski*, *schodytwarski*, *tepjenjetwarski*). Außerdem gibt es hier eine Reihe von partizipialen Bildungen, die grammatisch korrekten Syntagmen entsprechen und somit eigentlich getrennt zu schreiben wären (z. B. *krejzaprajacy* statt *krej zaprajacy*, *jenakmyslacy* statt *jenak myslacy*, *sobuorganizowany* statt *sobu organizowany*). Ähnliches gilt für Komposita mit den Zweitgliedern *hódny* und *połny* (*naspomnjenjahódny* statt *naspomnjenja hódny*, *oconapołny* statt *ocona połny*). Damit scheiden diese als Kandidaten für die Erweiterung der Lex-DB aus.

4.3.2 Niedersorbisch

Die aus den niedersorbischen Monitoring-Texten extrahierten bisher in der Lex-DB nicht verzeichneten Formen werden im Folgenden – wie zuvor schon für das Obersorbische – nach Wortarten gegliedert und auszugsweise vorgestellt. Da es in der gegenwärtigen frühen Projektphase auch um das Erproben von Analyseverfahren und Darstellungsformen für die Ergebnisse des Schrifttumsmonitorings geht, wird für das Niedersorbische versuchsweise und ergänzend zu den hier angeführten Beispielen eine komplette Liste

---

[108] Abweichungen von dieser Regel sind selten (*ameriskoruski*, *čěskomorawski*, *čornožołty*).

[109] So im PS 2005 bzw. 2014.

[110] Interessant ist, dass bei den registrierten Formen solche mit Erstglied ohne *n*-Suffix deutlich überwiegen (27:8, z. B. *wuchodobayerski*, *zapadobalkanski*, *zapadočěski* gegenüber *južnomorawski*, *wuchodnojuhowuchodny*), während es bisher in der os. Lex-DB ein ausgeglichenes Verhältnis gab (49:49). Ob sich hier ein Trend abzeichnet, lässt sich aufgrund der Datenmenge noch nicht sagen, zumal die meisten Belege aus „Serbske Nowiny“ stammen, die von einer einzelnen Person sprachlich endredigiert werden.

der vorläufig[111] als „neue Lexik“ eingestuften Formen im Internet bereitgestellt.[112] Die Liste ist in sich gegliedert, die einzelnen Kategorien neuer Lexik haben jeweils eine ID (z. B. [ID: K-01]), die an entsprechender Stelle, meist am Ende eines Abschnitts, angegeben wird.

In den folgenden Abschnitten (und in der Liste) werden jeweils Grundformen angeführt, denen im durchgesehenen Material jeweils mehrere Textbelege (in verschiedenen Flexionsformen) entsprechen können.

4.3.2.1 Verben

Nicht in der Lex-DB verzeichnet ist das Simplex *źaś* ‚sagen, sprechen‘. Dieses weist ein unvollständiges Paradigma auf. Da im Wörterbuch von Starosta (1999) mit dem Marker *alt* versehen, wurde das Verb zunächst bewusst nicht in die Datenbank aufgenommen.[113] Im Material des Monitorings tritt es vor allem – aber nicht ausschließlich – im religiösen Kontext auf, wobei es sich oft um Bibelzitate handelt.

4.3.2.1.1 Präfigierungen

Unter der bisher nicht registrierten Lexik findet sich eine Reihe von Präfigierungen, die in nicht seltenen Fällen eine Aktionsart ausdrücken. Dies ist nicht verwunderlich, da bisher weder alle im älteren Schrifttum belegten Bildungen in der Lex-DB erfasst sind, noch eine systematische Aufnahme aller potenziell möglichen Präfigierungen sinnvoll erscheint. Typische Beispiele sind etwa *dołamaś*, *nadundaś*, *pónygaś*, *zešnekotaś*.

Einige der beim Monitoring belegten Formen entsprechen u. U. nicht den Bestimmungen der aktuellen Kodifikation, z. B. *zwóttajaś*.[114] Es ist beachtenswert, dass entsprechende Beispiele gerade aus Texten muttersprachlicher guter Kenner der Standardsprache mit eher puristischer Einstellung stammen. Da weder das Aspektsystem noch die Bildung der Aktionsarten für das Niedersorbische ausreichend erforscht sind, stellen solche Formen für die weitere Erforschung des Niedersorbischen interessantes Material dar. [ID: K-01]

---

[111] Eine endgültige Bewertung kann erst zu einem späteren Zeitpunkt erfolgen (vgl. Kap. 5).

[112] Diese findet sich unter https://www.niedersorbisch.de/download/monitoring-2019-dsb.txt. Die Liste dokumentiert außer den als „neue Lexik“ eingestuften Formen in separaten Abschnitten auch nicht-standardsprachliches Material.

[113] Dies hängt mit der Enstehungsgeschichte der niedersorbischen Lex-DB zusammen, die seit 2014 zunächst ausschließlich als Grundlage für eine Applikation zur automatischen Rechtschreibprüfung erstellt wurde. Dieser Datenbestand wurde und wird aber mittlerweile – so auch im Rahmen des Schrifttumsmonitorings – schrittweise auch durch nicht synchron-normgerechte Lexik ergänzt. Vgl. Bartels 2020: Kap. 4.4, insbesondere die Fußnoten 50 und 51.

[114] Entsprechende Formen finden sich allerdings in Mucke (1911–1928). Sie scheinen jedoch in vielen Fällen systematisch, ohne Belegnachweise, gebildet worden zu sein. Zwar finden sich in der älteren Literatur entsprechende Beispiele, sie scheinen sich allerdings auf Ableitungen von wenigen Basismorphemen zu beschränken. Belegt ist beispielsweise der Typ *zwóttergaś*. Hier bedarf es entsprechender Untersuchungen.

4.3.2.1.2 Suffigierungen

Seltener finden sich im untersuchten Material bisher nicht registrierte Suffigierungen. Die entsprechenden Grundformen sind *nalicowaś*, *wóttšašowaś*,[115] *wutwórjowaś*. Bisher (und auch im untersuchten Material) nicht registriert sind auch die Verben *pśecynjowaś*, *póstajowaś*,[116] auf deren Grundlage die im Material belegten, bisher unverzeichneten Substantive *pśecynjowaŕ*, *samopóstajowanje*, *póstajowanje* gebildet wurden. Neben den genannten Verben ist auch desubstantivisches *farariś* (← *faraŕ*) belegt.

Nicht selten finden sich Fremdwörter, in denen das deutsche Suffix *-ieren* durch sorbisches *-ěrowaś* (*interesěrowaś*, *jubilěrowaś*), deutsches *-n* durch sorbisches *-owaś* (*pendlowaś*)[117] ersetzt wurde. Bei Verben mit fremdem Präfix und Wortstamm ist mit großer Wahrscheinlichkeit davon auszugehen, dass das Wort als ganzes übernommen wurde. Die Präfigierung hat also nicht innerhalb des Sorbischen stattgefunden und das Wort wurde mittels eines sorbischen Suffixes integriert. Ein Beispiel hierfür ist *dekompresěrowaś*. [ID: K-02]

4.3.2.1.3 Partizipien

Es gab verschiedene Gründe, im Material belegte Partizipien als „neu" zu klassifizieren. Einerseits handelt es sich um Fälle, in denen das Verb an sich bisher nicht verzeichnet ist. Dies trifft auf folgende Fälle zu: *interagěrujucy* (zu bisher unregistriertem *interagěrowaś*), *interesěrujucy* (zu *interesěrowaś*), *wupśedany* (zu *wupśedaś*),[118] *wusměwkujucy* (zu *wusměwkowaś*), *zajuskujucy* (zu *zajuskowaś*),[119] *zjadnośćony* (zu *zjadnośćiś*).[120] Auch *zwóttajany*, *zwóttergany* sind von Verben gebildet, die in aktuellen Wörterbüchern nicht verzeichnet sind. Allerdings wäre hier erst zu klären, inwieweit die Verben vom Typ *zwóttajaś* der schriftsprachlichen Norm entsprechen (vgl. Abschnitt Kap. 4.3.2.1.1 und Fußnote 114).

Zum anderen finden sich im Material Partizipien, die in der Lex-DB nicht registriert sind, obwohl das Verb an sich verzeichnet ist. In diesen Fällen handelt es sich um akzeptable Formen, die jedoch abweichend von den in der Datenbank registrierten Partizipien gebildet wurden: *wobspomnjety* (neben registriertem *wobspomnjony*), *wuschnuty* (neben *wuschnjony*), *pśimjety* (neben *pśimjony*), *pśispomnjety* (neben *pśispomnjony*). [ID: K-03]

4.3.2.2 Substantive

Die meisten der bisher nicht registrierten Substantive sind Komposita bzw. Derivate (s. u.). Jedoch sind auch andere Typen keine Seltenheit, meist handelt es sich dabei um Fremdwörter. Hierzu gehören zum einen Wörter wie *beamer*, *braindrain*, *loop*, *spray-*

---

[115] Das Verb *wóttšašowaś* ist in den aktuellen Wörterbüchern bisher nicht verzeichnet, findet sich aber bei Mucke (1911–1928).

[116] Es handelt sich dabei nicht um problematische Benennungslücken, da die imperfektiven Verben *pśecynjaś* und *póstajaś* verzeichnet sind. Allerdings sind Bildungen mit anderem Präfix und *cynjowaś* bzw. *stajowaś* lexikografisch belegt, bspw. *docynjowaś* und *wustajowaś*.

[117] Hier zusätzlich mit aussprachekonformem *e*-Ausfall.

[118] In den aktuellen Wörterbüchern finden sich allerdings *rozpśedaś*, *pśipśedaś*, *wótpśedaś*.

[119] In den aktuellen Wörterbüchern finden sich allerdings *pśijuskowaś* und *wobjuskowaś*.

[120] In Mucke (1911–1928) und Šwjela (1961) ist das Verb *zjadnośćiś* verzeichnet. Es kann aber nicht einfach als veraltet angesehen werden, sondern bedarf einer gesonderten Betrachtung.

er, die orthografisch (mit Ausnahme der Groß- bzw. Kleinschreibung) nicht angepasst wurden. In den meisten Fällen wurden sie jedoch problemlos in das Flexionssystem des Niedersorbischen integriert.

Ferner gibt es Fremdwörter, deren Schreibung an die sorbische Orthografie angepasst wurde, z. B. *awtograf*, *filipist*, *narratiw*, *propedeutikum*. Im Lehnwort *forwark*, das in den Wörterbüchern aus puristischen Gründen nicht enthalten ist, gibt es darüber hinaus lautliche Abweichungen gegenüber der heutigen, deutschen schriftsprachlichen Entsprechung *Vorwerk*. Morphologisch und orthografisch adaptiert sind u. a. *artroza*, *muzikolog/muzikologa* und *preparanda*. Fremdwörter werden oft auch durch reguläre Ersetzung von Wortelementen der Ausgangssprache (normalerweise Deutsch) durch im Sorbischen gebräuchliche Suffixe adaptiert. So wird bspw. im Entsprechungspaar *Biopsie – biopsija* deutsches *-ie* durch sorbisches *-ija* ersetzt. Weitere Beispiele für solche Ersetzungen werden vereinfachend zusammen mit den suffigierten Derivaten in Abschnitt 4.3.2.2.2 aufgezählt, obwohl es sich bei diesem Fremdwortadaptionstyp im Regelfall nicht um Derivation handelt (vgl. Pohončowa 2017: 73). Belegte Kurzwörter sind *app*, *elektro*, *reha*, *repro*. [ID: K-04]

4.3.2.2.1 Komposita

Im analysierten Material findet sich eine Vielzahl bisher nicht in der lexikalischen Datenbank erfasster Komposita unterschiedlichen Typs. Bei Bildungen mit Bindevokal sind ausschließlich Formen mit *-o-* belegt: *dalokostudij*, *krotkoportrej*, *nowopowědaŕ*, *samopóstajowanje*, *wjelikogmejna*, *wjelikomócnaŕstwo*, *wjelikowezir*. [ID: K-05]

Bei Bildungen ohne Bindevokal sind für den Typ mit Genitivform eines Substantivs im Erstglied (vgl. 4.2.1.2) *drastwynosaŕ*, *drastwynosaŕka*, *drastwyšlodaŕnja*, *rybow-łojenje* (korrekte Schreibung ohne Bindestrich), *rybow-łojśe* (korrekte Schreibung ohne Bindestrich) und für den Typ mit Genitivform eines Pronomens im Erstglied *sebjewuzgónjenje* belegt. Den Typ mit Akkusativform eines Substantivs im Erstglied belegen *groblucysćenje*, *jajamólaŕ*, *jajamólaŕka*, *jajamólowanje*, *kołoteptanje*, *wólejbiśe*. Als Komposita mit Adverb im Erstglied sind belegt *doma-byśe* (korrekte Schreibung ohne Bindestrich), *gromadupołoženje*, *dalejwuwijanje*, *sobugótowanje*, *sobuwandraŕ*, *zasejzałoženje*.[121] [ID: K-06]

Nicht selten treten Komposita auf, deren Erstglieder Abkürzungen, Akronyme, Einzelbuchstaben u. ä. darstellen. Ausgewählte Beispiele bisher nicht registrierter Formen dieses Typs sind *3D-śišć*, *AfD-šef*, *EU-projekt*, *LPG-pśedsedaŕ*, *SED-režim*. [ID: K-07]

Als Durchkopplungen sind belegt *pó-jaja-chójźenje*, *sam-se-póstajanje*, *do-šulechójźenje*, *acapella-spiwanje* (korrekte Schreibung mit Bindestrich auch im Erstglied: *a-capella-spiwanje*), *ad-hoc-pśestajenje*, *flora-fauna-habitat*. Als Zusammenrückungen stehen *pójsynoga* und *wšowěm* an der Grenze zwischen Komposition und Derivation. [ID: K-08]

---

[121] Die Erstglieder *dalej-*, *sobu-*, *zasej-* können auch als Präfixoide interpretiert werden (s. 4.3.1.2.2.3.1), *sobu-* wird in PS 2014 (für das Obersorbische) sogar explizit als Präfix geführt.

Recht häufig finden sich bisher nicht in der Lex-DB verzeichnete substantivische Hybridkomposita und zusammengesetzte, als Ganzes übernommene Fremdwörter.[122] Meist handelt es sich um Komposita mit Substantiv bzw. Adjektiv im Erstglied, vgl. *asana-zwucowanje*, *cover-kupka*, *cross-over*, *elektro-beat*, *elektromobilita*, *foto-workshop*, *gala-program*, *hip-hop*, *live-rozgrono*, *online-kurs*, *radijo-wusćełanje*, *rap-titel*, *seks-pupka*, *stasi-žłabnosć*, *šefkelnař*, *šlagerhit*, *Weddingplanařka* (korrekte Schreibung: *weddingplanařka*), *yoga-pozicija*. Innerhalb dieses Typs gibt es auch Komposita, deren Erstglied ein Kurzwort (Kopfwort) ist. Belegt sind *bio-mark*, *info-radijo*, *kombitiket*,[123] *org-běrow*, *popmuzikař*, *regio-bus*. [ID: K-09]

Bei den Komposita mit Namen im Erstglied dominieren Bildungen mit *Witaj*, vgl. *Witaj-camp*, *Witaj-ceptařka*, *Witaj-kubłanje*, *Witaj-mama*, *Witaj-wuknik*, *Witaj-źišownica* usw.[124] Außerdem findet sich *powerpoint-prezentacija* (korrekte Schreibung: *PowerPoint-prezentacija*). [ID: K-10]

4.3.2.2.2 Derivate

Der bisher nicht in der Lex-DB enthaltene Wortschatz umfasst eine Vielzahl an Wörtern, die durch Derivation entstanden sind. Im Bereich der Suffigierung fällt wie im Obersorbischen (vgl. Kap. 4.3.1.2.1.2) die nicht geringe Zahl von Deminutiven auf, da es sich um eine nach wie vor bei Muttersprachlern und Neusprechern mit gehobenem Sprachniveau sehr produktive Kategorie handelt und zudem auch etablierte Bildungen lexikografisch nicht vollumfänglich erfasst sind. In der Lex-DB fehlende Augmentativa wurden dagegen selten festgestellt.

Generell belegt sind Bildungen mit folgenden indigenen Suffixen: *-ař* (z. B. *blogař*, *kokotař*, *kšetař*, *wukazař*), *-ařka* (*móšynařka*, *wobsednikařka*, *zapisowařka*), *-ařnja* (*elektrařnja*), *-ica* (*rozgłosownica*), *-ick*/*-yck* (*jěžyck*), *-icka*/*-ycka* (*adwentnicka*, *pśestawcycka*, *žłařnicka*), *-ik*/*-yk* (z. B. *barik*, *dypcyk*, *karaselik*), *-ina* (*kšušwina*), *-isko* (*kokotisko* statt ebenfalls nicht verzeichnetem *kokośisko*, *žabisko*), *-išk*/*-yšk* (*cedlišk*), *-iško*/*-yško* (*kubliško*), *-išćo* (*žłařnišćo*), *-k* (z. B. *wagonk*), *-ka* (z. B. *anekdotka*, *elewka*, *kuchinka* [korrekt: *kuchyńka*],[125] *sekundka*, *SMSka*), *-ko* (*pšašańko*, *stojałko*), *-nik* (*AfD-nik,* weitere mögliche Schreibung *AfDnik*, *ansamblownik*), *-nja* (*budkařnja*, *filmařnja*), *-osć* (*cerkwinosć*, *drogotnosć*), *-owka* (*architektowka*, *promštowka*), *-ownja* (*gazownja*), *-stwo* (z. B. *direktorstwo*, *jadnařstwo*). [ID: K-11]

Bildungen mit nichtindigenen Suffixen sind vielfach nicht durch Wortbildungsprozesse innerhalb des Sorbischen entstanden, sondern sind das Ergebnis von Adaptions-

---

[122] Diese werden in den untersuchten Texten überwiegend mit Bindestrich geschrieben, in den Wörterbüchern herrscht hingegen die Zusammenschreibung vor. In den Interpunktionsregeln wird die Schreibung dieses Typs bisher nicht behandelt.

[123] Bildungen mit *kombi-* könnten auch als Derivate mit dem Präfix bzw. Präfixoid *kombi-* interpretiert werden.

[124] Nach dem Muster *Witaj-góle* gebildet sind auch *Lipa-góle*, *Lutki-góle*. Sie scheinen jedoch weitaus weniger konventionalisiert zu sein.

[125] In den aktuellen niedersorbischen Wörterbüchern ist potenziell zugrunde liegendes *kuchyn(j)a* nicht belegt. Mucke (1911–1928) bezeugt *kuchyńa* für den östlichen Grenz- (= Schleifer) und Muskauer Dialekt. Die Form *kuchinka* mit *-i-* ist wohl als Obersorabismus zu werten. Der betreffende Autor ist zwar einer der besten Kenner des Niedersorbischen, beherrscht allerdings auch das Obersorbische umfassend. In Mucke (1911–1928) ist die – zu erwartende – Form *kuchyńka* belegt.

prozessen, in denen oft im Deutschen verwendete Wortelemente durch ein im Sorbischen gebräuchliches Suffix ersetzt werden. So entspricht deutschem *-ation*/*-ierung* sorbisches *-acija* (z. B. *awtoprezentacija*, *deteritorializacija*), deutschem *-ie*/*-ion* sorbisches *-ija* (z. B. *ambrozija*, *biopsija*, *imersija*), deutschem *-ität* sorbisches *-ita*[126] (z. B. *intertekstualita*, *kreatiwita*). Für deutsches *-ismus* steht im Sorbischen standardsprachlich *-izm* (*hedonizm*), wärend die Beibehaltung von *-ismus* eher umgangssprachlich ist (*anticyganizmus* – korrekte Schreibung: *anticyganismus*). Bei deutschen Feminina auf *-ik* bzw. *-thek* wird *-ik* im Sorbischen durch *-ika* (z. B. *balkanistika*, *bulgaristika*, *sinfonika*) und *-thek* durch *-teka* (z. B. *mediateka*) ersetzt. [ID: K-12]

Verbalsubstantive werden von Fasske (1981: 331–335) als deverbale Substantive (und folglich nicht als paradigmatische Form des Verbs) betrachtet. Für die Lex-DB werden jedoch grundsätzlich zu sämtlichen registrierten Verben Verbalsubstantive gebildet, auch wenn dieselben nicht direkt in den zugrunde liegenden Wörterbüchern verzeichnet sind. Somit sind sämtliche Belege derselben eher als Belege für bisher nicht registrierte Verben zu betrachten. Es handelt sich dabei oft um Fremdwörter, teils aber auch um bisher nicht belegte Verbpräfigierungen,[127] vgl. *dosušenje*, *dowumrěše*, *koncertowanje*, *muzealizěrowanje*, *okanje* usw. [ID: K-13] Beispiele für Verbalsubstantive, die Komposita sind, finden sich in Abschnitt 4.3.2.2.1, zu Bildungen mit *dalej-*, *sobu-*, *zasej-* s. u.

Weitaus seltener als Suffigierungen finden sich unter den bisher nicht registrierten Wörtern Substantive, die durch ein Präfix erweitert wurden. Typisch sind in diesem Zusammenhang Negationen. Da diese systematisch gebildet werden können, werden sie in Wörterbüchern nur in speziellen Fällen separat geführt. Es ist also nicht verwunderlich, dass im untersuchten Material entsprechende Formen als „neu“ auftreten, z. B. *njelichota*, *njeměstno*, *njezachadnosć*. Das Wort *njekonsekwenca* ist die bisher nicht registrierte Alternative für *inkonsekwenca*.

Vielfach handelt es sich um Lehnübersetzungen aus dem Deutschen, vgl. Bildungen mit *mjazy-* (z. B. *mjazyfeedback*), *pód-* (*pódrubrika*) und *pśed-* (*pśedstudent*). Bildungen, die mit *dalej-*, *sobu-*, *zasej-* beginnen, werden hier als Komposita behandelt (vgl. 4.3.2.2.1), obwohl die entsprechenden Wortelemente auch als Präfixoide interpretiert werden können.[128]

Es treten auch Bildungen mit ursprünglich fremden Präfixen auf, z. B. *anti-* (*antigen*, *anti-wuglaŕ* – korrekte Schreibung ohne Bindestrich), *archi-* (*archidiakon*), *arcy-* (*arcymartraŕ*, *arcypresbyterium*), *eks-* (*ex-šołta* – korrekte Schreibung ohne Bindestrich), *poly-* (*polytechnika*), *pro-* (*pro-wuglaŕ* – korrekte Schreibung ohne Bindestrich), *super-* (*superfinale*, *superkral*). [ID: K-14] Belegt ist darüber hinaus die Kombination *super-mega-šula*, was wohl korrekt als *supermegašula* zu schreiben wäre.

---

126 In diesem Kontext ist auffällig, dass in den Wörterbüchern für deutsches *-ität* meist *-nosć* steht, obwohl *-ita* potenziell möglich wäre. In den für das Monitoring analysierten Texten findet sich Letzteres.

127 Bei einigen dieser präfigierten Verbformen wäre noch zu klären, inwieweit sie der schriftsprachlichen Norm entsprechen (vgl. Kap. 4.3.2.1.1 und Fußnote 114).

128 Im obersorbischen Rechtschreibwörterbuch PS 2014 wird *sobu-* sogar explizit den Präfixen zugerechnet. Starosta 1999 verzeichnet die Lemmata *dalej-*, *sobu-*, *zasej-*, die Wortart wird jedoch an entsprechender Stelle nicht spezifiziert.

4.3.2.3 Adjektive

In der Lex-DB bisher unberücksichtigt ist das synchron nicht abgeleitete (und etymologisch nicht ganz klare) *šwarny*. Dieses ist zwar im Wörterbuch in Starosta 1999 enthalten, dort aber als dialektal markiert und wurde deshalb zunächst nicht in die Datenbank übernommen (vgl. Fußnote 113). Als Fremdwort bisher in der Lex-DB ebenfalls nicht enthalten ist das indeklinable Adjektiv *mega*.[129] [ID: K-15]

4.3.2.3.1 Adjektivische Komposita

Einen nicht geringen Anteil der bisher nicht registrierten adjektivischen Komposita machen reihenbildende Formationen aus Zahlen bzw. Numeralien und Adjektiven aus. Potenziell ohne jede Einschränkung können Bildungen vom Typ Ziffer bzw. Numeral + *(-)lětny* auftreten. Im Material finden sich Beispiele für Schreibungen mit und ohne Bindestrich: *13-lětny*, *27-lětny*, *50-lětny*, *140-lětny*; *10lětny*, *14lětny*, *150lětny*, *1000lětny*. Für Zeitspannenangaben sind belegt *3-4lětny*, *57-lětny*. Seltener belegt sind Komposita dieses Typs mit dem Zweitglied *bocny* (*120-bocny*, Schreibungen ohne Bindestrich nicht belegt), *-minutowy* (*30minutowy*, Schreibungen mit Bindestrich sind nicht belegt), *-zwězkowy* (*10-zwězkowy*). Auch diese bilden potenziell unbegrenzte Reihen. Analoges gilt für Bildungen mit ausgeschriebenem Numeral, belegt sind *połnimski, połhundertlětny* und *pěśstowlětny*, *wósymbocny*, *wósymschójźeńkowy*, *dwaźasćalitrowy*, *towzyntgłosny*, *pěśzubaty*. Reihenbildend ist auch der Typ *10%ojski*, für den wohl die Schreibung *10%-ojski* zu präferieren wäre.[130]

Potenzielle Reihen bilden auch Farbangaben der Typen *carnožołty*, *carnomódry*, *carnosomośany*, *cerwjenomódry* (Mischfarben); *módro-běły*, *módro-cerwjeno-běły* (Zusammenstellung mehrerer Farben); *śmojtozeleny*. Dasselbe gilt für Adjektive, die die Beherrschung konkreter Sprachen zum Ausdruck bringen: *górnoserbskorěcny*, *pólskorěcny, rusojskorěcny*.

Nicht selten belegt sind zusammengesetzten Adjektive, die gegenseitige Beziehungen ausdrücken. Dies trifft vor allem auf Kombinationen von Völker- bzw. Sprachbezeichnungen zu, vgl. *nimsko-dolnoserbski*, *nimsko-francojski*, *nimsko-górnoserbski*, *nimsko-pólski*, *nimsko-rumuński*, *nimsko-słowjański; pólsko-česko-nimski*, *pólsko-nimski*, *pólsko-rusojski*, *pólsko-serbski*, *pólsko-słowakski; serbsko-dańsko-nimski*, *serbsko-nimsko-dański*, *serbsko-serbski* usw.

Auch Kardinal- und Ordinalzahlen – Fasske (1981: 339, 501) ordnet letztere den Adjektiven zu – sind in der Lex-DB bisher nur zu einem vergleichsweise geringen Teil enthalten. Eine vollständige Erfassung ist aufgrund der potenziell unendlichen Menge nicht möglich. Die in Wörterbüchern bisher eher sporadisch gebuchten Zahlenangaben könnten für die Lex-DB jedoch durch eine systematische Liste mit ausgeschriebenen

---

[129] Die konkreten Textbelege sind jedoch nicht eindeutig. Zumindest bei *MEGA kokota* (Großschreibung im Original) handelt es sich wahrscheinlich um fehlende Zusammenschreibung, *mega-* wäre in dem Fall ein Präfix. Im Textbeleg *MEGA party* könnte es sich tatsächlich um das Adjektiv *mega* im Sinne von ‚großartig, hervorragend' handeln. Auch dies ist jedoch auf Grundlage bisheriger Daten nicht klar ersichtlich.

[130] In den aktuell gültigen niedersorbischen Interpunktionsregeln wird dieser Fall nicht behandelt. Die entsprechende obersorbische Regel lässt ausschließlich die Schreibung mit Bindestrich zu.

Formen bis zu einer gewissen Schwelle ergänzt werden.[131] Im Material des Monitorings findet sich als bisher nicht belegtes Wort *źewjeśawósymźasety*. [ID: K-16]

4.3.2.3.1.1 Adjektivische Komposita mit Bindevokal

Viele der belegten adjektivischen Komposita sind, wie die bereits genannten Adjektive der Typen *nimsko-dolnoserbski*, *carnožołty*, *módro-běły* mit Bindevokal gebildet. Als solcher ist im bisher nicht registrierten Material des Monitorings ausschließlich *-o-* belegt. Besonders häufig vertreten sind adjektivische determinative *o*-Komposita, z. B. *českobratšojski*, *górnošlazyński*, *zakładnošulski.* Als kopulative adjektivische *o*-Komposita sind *analogisko-etymologiski*, *historisko-kritiski*, *kulturno-teritorialny* belegt.

Der Unterschied zwischen kopulativen (nach derzeitiger Norm in den meisten Fällen mit obligatorischer Bindestrichschreibung) und determinativen (derzeit obligatorische Zusammenschreibung) Komposita ist nicht immer ohne weiteres erkennbar. Außerdem scheinen die Interpunktionsregeln nicht für alle auftretenden Typen optimal zu sein. Es ist deshalb kaum verwunderlich, dass es viele Fälle unkorrekter Schreibungen gibt. Besonders häufen sich Schreibungen mit Bindestrich für determinative Komposita, z. B. *agrarno-strukturelny*, *stawiznisko-wědomnostny*. [ID: K-17]

4.3.2.3.1.2 Adjektivische Komposita ohne Bindevokal

Als bisher nicht verzeichnete adjektivische Komposita ohne Bindevokal sind neben den oben angeführten reihenbildenden belegt *cesćigłodny*, *rěcywucony* sowie die Zusammenbildung *wšoserbski*, *wšosokołski*. [ID: K-18]

4.3.2.3.2 Adjektivische Derivate

Als adjektivische[132] Derivate sind Bildungen mit verschiedenen Suffixen und Präfixen belegt. Suffigierungen [ID: K-19] sind gebildet mit *-any* (*interesěrowany*), *-aty* (z. B. *dujaty*, *kopcykaty*), *-iny*/*-yny* (z. B. *babcyny*), *-ny* (z. B. *ambrozijny*, *anglofony,*[133] *etnokulturny*), *-obny* (*lazujobny*), *-ojski* (*fabrikojski*), *-ojty* (*mrokawojty*), *-owy* (z. B. *póglědnicowy*, *WGejowy*), *-ski*[134] (z. B. *dendrochronologiski*, *ewaluaciski*, *monotematiski*)[135].

Das Suffix *-ny* tritt öfter auch beim Adaptionsprozess von adjektivischen Fremdwörtern auf, vgl. *dement – dementny*, *fiktional – fikcionalny*, *transkulturell – transkulturelny*. Gleiches gilt für *-arny* in *ewolucionarny*.

---

131 In der os. Lex-DB sind die ausgeschriebenen Kardinalien und Ordinalien bis 99 systematisch ergänzt worden. Zumindest für die Kardinalien ist dies ausreichend, da höhere Zahlenwerte repräsentierende Numeralien (bis auf die in der Lex-DB ohnehin enthaltenen ganzen Hunderter) getrennt geschrieben werden.

132 Vereinfachend werden hier die Possesiva gemeinsam mit den Adjektiven behandelt, vgl. jedoch Fasske (1981: 381–385).

133 Das Wort *anglofony* ist wohl als Bildung mit *-ny* zu ausgangssprachlichem *anglofon* zu interpretieren, bei dem entstandenes *-nn-* rechtschreiblich zu *-n-* vereinfacht wurde.

134 Für die Bildung von Adjektiven, die von geografischen Namen abgeleitet sind, werden oft die zusammengesetzten Suffixe *-ański* bzw. *-ojski* genutzt, vgl. *cerkwicański*, *rujański*, *škrjokojski*.

135 Die Formen *dendrochronologiski* und *monotematiski* sind wohl nicht als eigenständige sorbische Bildungen zu *chronologiski* bzw. *tematiski* mit den Präfixen *dendro-* bzw. *mono-* zu werten. Es ist davon auszugehen, dass die Ableitungsbasis insgesamt, also bereits präfigiert, aus dem Deutschen übernommen wurde.

Potenziell unbegrenzt reihenbildend sind die herkömmlich als Ordinalia bezeichneten Adjektive. In den Texten ist vor allem die Schreibung derselben vom Typ Ziffer + *-ty* belegt: *16ty*, *1870ty*.

In großer Zahl treten von Namen gebildete Possessiva auf *-owy* bzw. *-iny*/*-yny* auf, z. B. *Pankowy*, *Mininy*. Im Unterschied zum Obersorbischen treten im Niedersorbischen auch von Namen, die auf männliche Personen verweisen, mit *-iny*/*-yny* gebildete Formen auf: *Jozuwiny.* [ID: K-20]

Von einer Präpositionalphrase mit Hilfe von *-ny* abgeleitet wurde *mjazykulturny* (aus *mjazy kulturoma/kulturami*), ein deverbales Adjektiv ist *pśedcytański* zu *pśedcytaś*.

Präfigierungen finden sich bei den deadjektivischen Adjektiven. Erwartungsgemäß ist das Negationsaffix *nje-* das meistgenutzte Präfix in diesem Kontext, vgl. *njedaloki*, *njekomercielny*, *njeśwarny*. Als weitere Präfixe treten auf *pó-* (*póslobrany*), *pśed-* (z. B. *pśedsłowjański*) und *wob-* (*wobsuchy*). Auch das systematisch zur Bildung von Gradationsformen genutzte Präfix *pśe-* ist mehrfach belegt (z. B. *pśenjerealistiski*). [ID: K-21]

4.3.2.4 Neuentlehnungen

Es sind Neuentlehnungen belegt, für die in der niedersorbischen Schriftsprache bereits ältere Entlehnungen etabliert sind. Die neuen Formen füllen also keine Benennungslücken. Konkret sind zu nennen *centnaŕ* statt *cantnaŕ*, *tona* statt *tuna* sowie *tafla*, *taflicka* statt *tofla*, *toflicka*. In allen Fällen ist die neue Entlehnung an die heutige deutsche Standardsprache angelehnt, während die etablierten Formen auf (älteren) dialektalen deutschen Formen basieren.

4.3.2.5 Impulse für die Schließung von Benennungslücken trotz offensichtlicher Fehler bzw. missglückter Bildungen

Obwohl einige bisher nicht registrierte Formen offensichtlich Fehler sind, können sie trotzdem Benennungslücken sichtbar machen und so Impulse zum lexikalischen Ausbau des Niedersorbischen geben. So ist zwar *FDlaŕ* als Lapsus zu werten, allerdings befindet sich intendiertes *FDJlaŕ* tatsächlich bisher nicht im Datenbestand. Ebenso entspricht zwar die Schreibung *samarijaŕ* nicht den Regeln der orthografischen Fremdwortadaption, tatsächlich wäre aber eine Unterscheidung zwischen *Samariaŕ*, *Samariski* (‚Bewohner von Samarien') und *samaritaŕ/Samaritaŕ* (‚Samariter') bzw. *Samaritanaŕ* (‚Samaritaner', ebenfalls bisher nicht registriert) durchaus sinnvoll. Auch *wiolonist* wird zwar im konkreten Kontext falsch statt lexikografisch registriertem *wiolinist* verwendet. Das Wort ist aber durchaus in der Bedeutung ‚Person, die Violone spielt' denkbar. Weitere Beispiele für Fehlschreibungen bzw. -bildungen, die Impulse für den lexikalischen Ausbau des Niedersorbischen geben, sind *internet-platform* → *internetowa/internetna platforma*, *rakečański* → *rakecański*, *profesionalzěrowanje* → *profesionalizěrowanje*.

4.3.2.6 Obersorabismen

Historisch und/oder biografisch bedingt ist bei einem nicht unbeträchtlichen Teil der heute niedersorbisch Schreibenden ein mehr oder weniger starker Einfluss des Obersorbischen festzustellen. Im untersuchten Material wurde daher eine beträchtliche Anzahl lexikalischer Obersorabismen festgestellt. In den meisten Fällen sind dafür in den

zugänglichen relevanten Wörterbüchern formal abweichende niedersorbische Entsprechungen enthalten. Teilweise handelt es sich um phonetische (z. B. *byrgański* statt *bergański*) oder morphologische Abweichungen (z. B. *dnjowy* statt *dnjowny*, *kaznje* – als Nom. Pl. – statt *kazni*), Unterschiede in der Wortbildung (z. B. *kowańja* statt *kowalnja*) oder um völlig verschiedene Lexeme (z. B. *tołmač* statt *dolmetšaŕ*, *tonuška* statt *ryborak*). [ID: K-22]

Normgerecht hingegen ist, wenn obersorbische Eigennamen wie im Obersorbischen geschrieben werden und in den abhängigen Fällen die niedersorbischen Endungen unter Berücksichtigung des entsprechenden Konsonantenwechsels im Stammauslaut zur Anwendung kommen, z. B. *Hornja Kina*. Wenn der Nominativ des Eigennamens *Zejler* im Obersorbischen mit *r* endet, so lauten die deklinierten Formen nach derzeitiger Norm *Zejlera*, *Zejleroju*, *Zejlerom*, *Zejlerje/Zejleru*. Dasselbe gilt für Ableitungen: *Zejlerowy*. Im Material mehrfach belegte Formen vom Typ *Zejlerja*, *Zejlerjowy* entsprechen nicht dem heutigen Stand der Norm. Es wäre zu überlegen, ob bei gewissen Namentypen nicht eine Adaptierung auch im Nominativ (Typ *Zejleŕ*, mit dann entsprechend deklinierten Formen *Zejlerja* usw.) sinnvoll wäre.

4.3.2.7 Impulse für mögliche Änderungen der Kodifizierung

In einigen Fällen könnten bisher nicht registrierte Formen Anlass für Änderungen der standardsprachlichen Norm geben. Relativ häufig trifft das auf deklinierte Formen von Fremdwörtern zu, die im derzeitigen Datenbestand als indeklinabel geführt sind. Mit zunehmender Nutzung der betreffenden Wörter steigt offenbar der Druck zur Integration ins Sprachsystem. Betroffen sind u. a. *action*, *duo*, *e-mail*, *echo*, *piano* (Instrument).

In einigen Fällen scheint die Zuordnung von Varianten zum Standard bzw. den Dialekten eher zufällig getroffen worden zu sein. Ein Beispiel hierfür wäre *nĕgajšy*, das im traditionellen Schrifttum in etwa mit gleicher Frequenz wie das dem aktuellen Standard entsprechende *nĕgajšny* auftritt.[136] Da die niedersorbische Standardsprache allgemein Varianten verschiedenen dialektalen Ursprungs bewusst akzeptiert (vgl. *tykańc* vs. *mazańc*, *kupiju* vs. *kupijom* usw.), wäre auch hier zu prüfen, ob die (erneute) Aufnahme von *nĕgajšy* in den Standard sinnvoll wäre.

Interessant sind die belegten Formen des Wortes *kralejstwo* mit *-ej-* statt kodifiziertem *-oj-*. Sie stammen wohl alle aus Zitaten aus dem älteren Schrifttum. Tatsächlich steht die heutige Norm dem älteren Usus beinahe diametral gegenüber.[137] Allerdings könnte sich hier ein Sprachwandel vollzogen haben.[138]

Auch die Adaptionsregeln für (einige Typen) obersorbischer Eigennamen könnten überdacht werden, vgl. 4.3.2.6.

---

[136] In der Bibelausgabe von 1868 bspw. wird der Typ *nĕgajšy* (18mal) gegenüber *nĕgajšny* (10mal) bevorzugt.

[137] Etwa 95 Prozent der relevanten Belege im Online-Korpus Dotko, das Texte bis 1937 zugänglich macht, weisen die Schreibungen *-ej-*, *-ei-*, *-ey-* auf. Gesucht wurde nach /krale[ijy]ſt[wẃẇ].*/ (3418 Treffer) bzw. /kralo[ijy]ſt[wẃẇ].*/ (187 Treffer).

[138] Darauf könnte hindeuten, dass Mucke (1911–1928) die heute kodifizierte Norm in sein Wörterbuch aufnimmt und als umgangssprachlich klassifiziert. Das Wörterbuch von Starosta (1999) versieht die Form *kralejstwo* dann mit dem Qualifikator *alt* und verweist auf *kralojstwo*.

4.4. Nutzen des Schrifttumsmonitorings für Sprachbeschreibung, -kodifizierung und -unterricht

Die vorgestellten vorläufigen Ergebnisse der Analyse des aktuellen sorbischen Schrifttums belegen neben den primär angestrebten Nutzungsmöglichkeiten für Sprachdokumentation, Sprachwandelforschung und lexikografische Aufgaben erwartungsgemäß auch ein großes Potenzial mit Blick auf die synchrone Beschreibung der Sprachsysteme des Ober- und Niedersorbischen (einschließlich eines stellenweise erkennbaren Forschungsbedarfs) sowie für die Sprachkodifizierung und den Sprachunterricht.

Wie explizit die vorangegangenen Abschnitte 4.3.2.6 und 4.3.2.7, so enthalten auch schon die beiden Kapitel 4.2 und 4.3 zahlreiche Hinweise auf die Notwendigkeit von Anpassungen der aktuellen Sprachbeschreibung oder zur möglicherweise sinnvollen Veränderung der geltenden Normierungen. Teilweise deuten häufige und relativ systematische Fehler in den Texten aber auch auf Defizite in der oder auf Aufgaben für die Sprachausbildung hin. Für Letzteres soll hier noch ein Beispiel aus dem Niedersorbischen gegeben werden.

In den beim Monitoring berücksichtigten Texten treten stellenweise Formen perfektiver Verben der ursprünglichen *i*- und *a*-Konjugation auf, die nicht mit Hilfe des Infixes *-jo-* in die *o*-Konjugation überführt worden sind. Diese (nicht erweiterten) Formen gelten heute als veraltet. Sie müssen wohl mehrheitlich als Relikte der Zeit nach 1945 gewertet werden, als die typischen, bereits in den ältesten niedersorbischen Texten belegten, mit *-jo-* erweiterten Formen unter dem Einfluss des Obersorbischen aus der Schul- und weitgehend auch aus der damaligen Standardsprache verdrängt wurden, was eine auffällige Differenz zur Volkssprache verursachte. Erst gegen Ende der DDR-Zeit sind die traditionellen Formen in schriftlichen Texten wieder in größerem Umfang in Gebrauch gekommen, spätestens gegen Ende der 1990er-Jahre haben sie sich wieder als die Hauptformen durchgesetzt. Viele der heute schreibenden Autoren haben das Niedersorbische jedoch zur Dominanzzeit der nicht erweiterten Formen erlernt und verwenden diese (bisweilen) noch heute.

Insgesamt treten in den untersuchten Texten 67 solcher Formen auf, die meisten entfielen auf die Verben *zmakaś* (8), *namakaś* (6), *póglědaś* (6)[139] sowie *pótrjebaś* (6)[140]. Für die Verben *pópowědaś*, *powdaś se*, *pśeměniś*, *pśibliżyś*, *wózjawiś*, *wrośiś*, *wupóraś*, *zaspiwaś*, *znicyś* sind jeweils zwei Beispiele mit veralteter Konjugationsform belegt. Einzelbelege gibt es für die Verben *łapiś*, *napóraś*, *póśćiś*, *póstaraś*, *pśemjelcaś*, *pśistajiś*, *pśiškriniś*, *pśiwabiś*, *starcyś*, *wobźiwaś*, *wóstajiś*, *wótleśeś*, *wuchójźiś*, *wuleśeś*, *wurownaś*, *wuwitaś*, *zabincaś*, *zachopiś*, *zakyrcaś*, *zasłyšaś*, *zdźaržaś*, *zgóniś*, *zraniś*. Die Mehrzahl der Belege (49) stammt aus dem Nowy Casnik (Ns-1).[141]

---

[139] Es handelt sich jeweils um die Form *Póglědamy*, die als Titel innerhalb einer sich in mehreren Ausgaben des Wochenblatts wiederholenden Rubrik mehrmals verwendet wird.

[140] In allen Fällen könnte bzw. müsste die präfigierte und somit perfektive Form durch nicht präfigiertes, imperfektives *trjebaś* ersetzt werden.

[141] Weitere acht Konjugationsformen dieses Typs stammen aus Ns-6, jeweils vier Belege mit nicht erweiterten Formen finden sich in der „Serbska Pratyja“ (Ns-7) und in niedersorbischen Texten des „Rozhlad“ (Ns-2), weitere zwei in Ns-8.

Umgekehrt finden sich in den untersuchten Texten auch 16 sichere Belege für hyperkorrektes Übertragen der *jo*-Erweiterung auf imperfektive Verben.[142] Bis auf einen stammen alle aus Ns-1. Betroffen sind – in Anlehnung an die perfektiven Verben vom Typ *kupiś* – vor allem imperfektive Simplizia der *i*-Konjugation: *basliś* (1), *jawiś* (2), *leśeś* (1), *mokśiś* (1), *pisaniś* (1), *płomjeniś* (1), *pyšniś* (1), *sławiś* (1), *sušyś* (1), *wjaseliś* (2). Seltener betroffen sind Verben der *a*-Konjugation: *cytaś* (1), *glědaś* (1), *wobsajźaś* (1), *wótchylaś* (1).[143]

4.5. Möglichkeiten sprachstatistischer Aussagen und Darstellungen

Im vorliegenden ersten Monitoring-Bericht finden sich natürlich Belegzahlen wie auch Prozentangaben zur absoluten oder relativen Häufigkeit bestimmter Formen oder Formkategorien. Das Diagramm in Kap. 3.2.2.2.3.1 bietet eine einfache Visualisierung zu quantitativen Verhältnissen bei „nicht-niedersorbischer Lexik“ in den analysierten ns. Korpustexten. Wie bereits in Bartels 2020: Kap. 4.6 erwähnt, wird die für das Schrifttumsmonitoring betriebene Aufbereitung von Korpustexten mit wachsender Datenbasis zunehmend die Möglichkeit eröffnen, das Schrifttum auch statistisch auszuwerten, Entwicklungen zu erkennen, zu visualisieren und auf verschiedene Weise zu interpretieren. Eine naheliegende Methode ist z. B die sog. „keyword analysis“ (Gabrieletos 2018), die verschiedene Formen lexikalischer Dynamik im Schrifttum sichtbar machen kann. Diese Methode einzusetzen ist aber derzeit noch nicht sinnvoll, denn: „Keyword analysis does not make visible what is characteristic of a set of data, but what is characteristically different between that set and another set“ (Marchi 2018: 181 f.). Es geht hier stets um einen mit statistischen Verfahren unternommenen Frequenzvergleich zwischen verschiedenen Korpora oder angemessen definierten Datenmengen oder auch zu einer zuvor durch vorgeschaltete Analysen festgestellten Norm. Wie bereits in Kap. 2.3 mit Blick auf die Ermittlung von Neologismen beschrieben, dient das Schrifttumsmonitoring auch dazu, durch Schaffung einer validen Datengrundlage einen entsprechenden Einsatz statistischer Verfahren zu ermöglichen. Wir hoffen, bereits nach einigen Jahren des Monitorings erste derartige Analysen liefern zu können.

---

[142] Berücksichtigung finden hier auch Formen der 3. Pers. Pl. imperfektiver Verben der *i*-Konjugation, die in Analogie zum perfektiven Simplex *kupiś* auf *-iju* (statt auf *-’e*) auslauten.

[143] Unklar sind hingegen vier belegte Formen vom Typ *zdajo se* (statt *zda se*, zu *zdaś se*). Sie könnten potenziell als weitere Belege für allgemeines hyperkorrektes Verwenden der *jo*-Erweiterung bei imperfektiven Verben der *a*-Konjugation (oder als Anlehnung an perfektives *daś*) gewertet werden. Möglich (und wahrscheinlich) ist aber auch Analogie zu den einsilbigen imperfektiven Verben vom Typ *graś*. Die Formen *grajom*, *grajoš* sind nicht durch sekundäre Erweiterung mit Hilfe des Infixes *-jo-* entstanden, sondern sie kontinuieren alte, unkontrahierte Formen. – Auch die Form *łapajo* zu imperfektiven *łapaś* kann auf zweierlei Weise interpretiert werden: Zum einen als hyperkorrekte Form mit *-jo*-Erweiterung, zum anderen als Kontamination der beiden möglichen Formen *łapjo* (nach der *o*-Konjugation) und *łapa* (nach der *a*-Konjugation).

5. Bilanz des ersten Monitoring-Jahres und Ausblick

Beim vorliegenden Text handelt es sich um den ersten Bericht aus einem neuen und längerfristig angelegten Forschungsvorhaben und damit um einen Einblick in Arbeit und Zwischenergebnisse einer frühen Projektphase. Die auf zwei Jahre (2019/20) angelegte Pilotphase ist bei Erscheinen dieses Artikels im Frühjahr 2021 zwar formal beendet. Aus dem dargestellten Verfahren und der sich daraus logisch ergebenen zeitlichen Abläufe sollte aber deutlich geworden sein, dass diese erste Phase inhaltlich erst 2022, nach der Analyse des zuvor aufbereiteten digitalen Schrifttums aus dem Jahre 2020 (im Laufe des Jahres 2021) und der anschließenden Ergebnisverwertung (Bericht, Eingliederung in die Datenbasen für die automatische Rechtschreibkontrolle usw.), abgeschlossen werden kann.

Währenddessen hat eine dreijährige „Konsolidierungsphase" des Vorhabens (2021–23) begonnen, die sich entsprechend inhaltlich bis 2025 erstrecken wird. Was ist in den kommenden Jahren zu konsolidieren? In erster Linie geht es um eine Weiterentwicklung des Verfahrens und der eingesetzten Werkzeuge. Während das Verfahren zur Herstellung hochwertiger Korpustexte – ein wichtiges Teilziel des Monitorings – bereits in den Jahren zuvor entwickelt worden war und nun nur noch (z. B. mit Blick auf die Formate der vom LND gelieferten Texte) angepasst werden musste, wurde die geschilderte Analyse und Auswertung des Sprachmaterials im vollen Umfang erstmals durchgeführt. Der Gesamtaufwand ist hoch und muss auch mit Blick auf die Verfügbarkeit von Projektmitteln wie auch entsprechend qualifizierten Personals reduziert werden – die schrittweise Anreicherung der beiden lexikalischen Datenbanken ist hierfür das wichtigste Instrument. Dabei müssen in den nächsten Jahren wichtige Lücken (wie derzeit beim Obersorbischen die Ergänzung bisher nicht erfasster Lexik aus dem DOW 1989/91 oder eine Inventarisierung und Einbeziehung von Eigennamen) geschlossen werden. Ebenso sollten Beschränkungen der derzeitigen Korpustext-Annotation (vgl. Bartels 2020: Kap. 4.4.7) mittelfristig aufgehoben werden. Auch die dargestellten Unterschiede im Vorgehen bei der Analyse der niedersorbischen bzw. obersorbischen Texte sollen in Zukunft möglichst weitgehend verringert werden, zumindest in dem Maße, wie sie sich auf die Art der gewonnenen Daten und damit unter Umständen auf die Ergebnisse und deren Präsentation auswirken. Das mit Blick auf Erkenntnisziele gemeinsame Vorgehen und die parallele Betrachtung beider sorbischer Schriftsprachen ermöglicht jedoch eine Schärfung der kontrastiven Perspektive und wird Impulse für die Weiterentwicklung auch von Beschreibungskategorien in der sorabistischen Grammatikschreibung liefern. Und schließlich ist zu überlegen, um welche Texte das bisher zugängliche Schrifttum, das ja zunächst ausschließlich aus Publikationen aus dem Domowina-Verlag besteht, ergänzt werden sollte – Kap. 3.2.2 liefert hier mit dem Verweis auf die konfessionellen Zeitschriften „Pomhaj Bóh" und „Katolski Posoł" bereits wichtige Handlungsempfehlungen.

Abgesehen von den kurzfristigen und spezifischen Nutzungen der Ergebnisse des Schrifttumsmonitorings (z. B. der Einbindung neuer Lexik in die Applikationen für die automatische Rechtschreibkontrolle sowie die laufende Erarbeitung hochwertiger Korpustexte) und seinen langfristigen Zielen (v. a. dem Aufbau einer zuverlässigen Datenbasis für Lexikografie und Grammatikschreibung wie auch für Untersuchungen zur Sprachentwicklung) ist schon in diesem ersten Jahresbericht erkennbar, dass aus dem Projekt auch wertvolle Hinweise für Sprachausbildung und Sprachkodifizierung hervorgehen (vgl. Kap. 4.4). So werden Lücken in der aktuellen Sprachbeschreibung ebenso identifiziert wie Probleme der geltenden Norm. Diese können als Nebenprodukte des

Schrifttumsmonitorings direkt in normsetzende Publikationen (Rechtschreibwörterbücher, Regelwerke) eingehen oder – verbunden mit Lösungsvorschlägen – an die entsprechenden Zielgruppen vermittelt werden (z. B. über Artikel in der Fachzeitschrift „Serbska šula“).

Darüber hinaus geht es in nächster Zeit auch darum, nach Wegen zu suchen, wie die identifizierte „neue Lexik“ möglichst schnell und gut nutzbar öffentlich zur Verfügung gestellt werden kann, ohne dass damit vorschnell „halbgare“ Lösungen in die Welt gesetzt werden. Mehrfach wurde darauf hingewiesen, dass in den Texten belegte Formen nicht einfach zu bewerten sind, dass in manchen Fällen die Frage der Korrektheit bzw. Angemessenheit nicht einfach zu beantworten ist, dass sich gelegentlich auch die Frage stellt, ob nicht eher die geltende Norm angepasst werden sollte. Derartige Fragen mitsamt entsprechender Belege klar vor Augen zu haben, war ein Ziel des Monitorings. Antworten liegen in einigen Fällen klar auf der Hand, verlangen aber nicht selten eine genauere Prüfung oder die Einbeziehung der beiden Sprachkommissionen; in manchen Fällen ist es sogar notwendig, zunächst Daten aus weiteren Monitoring-Jahren zu sammeln. Die klaren Fälle geeigneter „neuer Lexik“ sollen aber zügig über die beiden Sprachportale niedersorbisch.de und obersorbisch.de (unter Einbeziehung von soblex.de) zugänglich gemacht werden.

Bei allen geschilderten Herausforderungen, die mit der Verfolgung der im Projekt gesteckten Ziele (vgl. Kap 1 und 2) einhergehen, bietet das Schrifttumsmonitoring, sofern es wie geplant längerfristig umgesetzt wird, ein enormes Potenzial für die sorabistische Sprachwissenschaft, insbesondere mit Blick auf eine synchrone und aktuelle Beschreibung der beiden modernen sorbischen Schriftsprachen, aber auch mit Blick auf ihr „Gewordensein“ und aktuelle wie rezente Wandelerscheinungen. Es entsteht eine valide Datenbasis für die zeitgenössische sorbische Lexikografie. Und sofern es gelingt, auch ein rückwärtiges Monitoring zu ergänzen: auch für eine moderne historische Wörterbuchschreibung, auf deren Grundlage eine endgültige Bewertung, welche in aktuellen Texten vorkommende Lexik tatsächlich (und ggf. auf welche Weise) „neu“ sei, erst ermöglicht wird (vgl. Kap. 2.3). Die langfristige Bedeutung des neuen Schrifttumsmonitorings für die Sorabistik sollte daher nicht unterschätzt werden.

Literatur

Bär, Jochen A. 2020: Was im neuesten Duden fehlt. Ein sentimentalischer Streifzug, in: Graf 2020, S. 223–232.

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 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. 2013, 37–46. [Niedersorbische Fassung: Ku koncepciji historisko-dokumentěrujucego informaciskego systema za dolnoserbski słowoskład. Plany k wótpóranju nuznego slěźeńskego deziderata, in: Lětopis 60/1, S. 16–26.]

Bartels, Hauke 2020: Das niedersorbische Globalkorpus als Ziel einer ganzheitlichen Konzeption zum Aufbau von Textkorpora, in: Lětopis 67/2, S. 3–44.

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/ [22.01.2021].

Dotko: Dolnoserbski tekstowy korpus, Internet:
(a) https://dolnoserbski.de/korpus/standard/pytanje [22.01.2021];
(b) https://www.korpus.cz/kontext/query?corpname=dotko [22.01.2021].

DOW 1989/1991 – Jentsch, Helmut; Michalk, Siegfried; Šěrak, Irene: Deutsch-obersorbisches Wörterbuch, Band 1 A–K. Bautzen 1989, Band 2 L–Z. Bautzen 1991.

DOW-NL 2006 – Jentsch, Helmut; Pohontsch, Anja; Schulz, Jana: Deutsch-obersorbisches Wörterbuch neuer Lexik. Bautzen.

Engelberg, Stefan; Lemnitzer, Lothar 2009: Lexikographie und Wörterbuchbenutzung. 4., überarbeitete und erweiterte Auflage. Tübingen.

Exonyme: Datenbank obersorbischer geografischer Exonyme, Internet: https://www.serbski-institut.de/os/Geografiske-mjena-hornjoserbsce/ [22.01.2021].

Faska, Helmut 1998 (Hg.): Serbšćina. Najnowsze dzieje języków słowiańskich. Opole.

Faska, Helmut 2012: Pućnik po hornjoserbšćinje. Gramatika. Budyšin.

Fasske, Helmut 1981 (unter Mitarbeit von Siegfried Michalk): Grammatik der obersorbischen Schriftsprache der Gegenwart. Morphologie. Bautzen.

Gabrieletos, Costas 2018: Keyness Analysis: Nature, Metrics and Techniques, in: Taylor/Marchi 2018, S. 225–258.

Graf, Peter 2020: Was nicht mehr im Duden steht. Eine Sprach- und Kulturgeschichte. Berlin.

Herberg, Dieter; Kinne, Michael; Steffens, Doris 2004: Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen. Berlin-New York.

Hotko: Hornjoserbski tekstowy korpus, Internet: https://www.korpus.cz/kontext/query?corpname=hotko [22.01.2021].

HRK 2007: Nowe postajenja na polu ortografije, morfologije a interpunkcije w hornjoserbšćinje wobzamknjene wot Hornjoserbskeje rěčneje komisije w lětomaj 2002 a 2007, Internet: https://www.domowina.de/fileadmin/Assets/Domowina/Material_MacicaSerbska/Material_Macica_Dokumente/Nowe_prawidla.pdf [22.01.2021].

Jakubaš, Filip 1954: Hornjoserbsko-němski słownik, Obersorbisch-deutsches Wörterbuch. Budyšin.

Janaš, Pětr 1984: Niedersorbische Grammatik für den Schulgebrauch, 2., durchgesehene Auflage. Bautzen.

Jentsch, Helmut 1999: Die Entwicklung der Lexik der obersorbischen Schriftsprache vom 18. Jahrhundert bis zum Beginn des 20. Jahrhunderts. Bautzen (= Schriften des Sorbischen Instituts; 22).

Kaulfürst, Fabian 2019: Pśinosk k dolnoserbskej ortoepiji na zakładze projekta awdijowych datajow za nimsko-dolnoserbski internetowy słownik, in: Lětopis 66/1, S. 3–41.

Kinne, Michael 1998: Der lange Weg zum deutschen Neologismenwörterbuch, in: Teubert Wolfgang (Hg.) 1998: Neologie und Korpus. Tübingen, S. 63–110.

Kral, Georg 1927: Wendisch-deutsches Wörterbuch der ober-lausitzer Sprache. Bautzen.

Marchi, Anna 2018: Dividing up the Data. Epistemological, Methodological and Practical Impact of Diachronic Segmentation, in: Taylor/Marchi 2018, S. 174–196.
Meškank, Timo 2017: Serbske předmjena. Serbske pśedmjenja. Sorbische/wendische Vornamen. Budyšin.
Michałk, Frido 1974: Slovoobrazovanie, in: Trofimowič, Konstatin K. (Hg.): Hornjoserbsko-ruski słownik. Budyšin-Moskwa 1974.
Mucke, Ernst 1911–1928: Wörterbuch der nieder-wendischen Sprache und ihrer Dialekte. St. Petersburg.
Pfuhl, Christian Traugott 1866: Łužiski serbski słownik. Lausitzisch Wendisches Wörterbuch. Budissin.
Pohončowa, Anja 2017: Zarys hornjoserbskeje słowotwórby, in: Lětopis 64/1, S. 71–86.
PS 1981: Völkel, Pawoł: Hornjoserbsko-němski słownik. Obersorbisch-deutsches Wörterbuch. Prawopisny słownik hornjoserbskeje rěče. 4., sylnje předźěłany a rozšěrjeny nakład. Budyšin.
PS 2005: Völkel, Pawoł: Prawopisny słownik hornjoserbskeje rěče, wobdźěłał Timo Meškank. 5., wobdźěłany a sylnje rozšěrjeny nakład. Bautzen.
PS 2014: Völkel, Pawoł: Prawopisny słownik hornjoserbskeje rěče, wobdźěłał Timo Meškank. 6., přehladany nakład. Bautzen.
Rechtschreibduden 2020: Duden. Die deutsche Rechtschreibung. 28., völlig neu bearbeitete und erweiterte Auflage. Hrsg. von der Dudenredaktion. Berlin.
Rěčne kućiki: Pohončowa, Anja; Šołćina, Jana; Wölkowa, Sonja: Manuskripte von sprachkulturellen Beiträgen unter https://hornjoserbsce.de/kuciki/ [22.01.2021].
Rězak, Filip 1920: Němsko-serbski wšowědny słownik hornjołužiskeje rěče. Deutsch-wendisches encyklopädisches Wörterbuch der oberlausitzer Sprache. Bautzen.
Scholze, Lenka 2008: Das grammatische System der obersorbischen Umgangssprache im Sprachkontakt. Bautzen (= Schriften des Sorbischen Instituts; 45).
SSA – Sorbischer Sprachatlas. Bde. 1–10, bearb. von Helmut Fasske, Helmut Jentsch und Siegfried Michalk. Bautzen 1965–1986; Bd. 11: Morphologie. Die grammatischen Kategorien. Die Paradigmatik des Substantivs, bearb. von Helmut Fasske. Bautzen 1975; Bd. 12: Morphologie. Die Flexion der Adjektive, Pronomen und Verben, bearb. von Helmut Fasske. Bautzen 1988.
Starosta, Manfred 1992: Niedersorbisch schnell und intensiv, Bd. 2. Bautzen.
Starosta, Manfred 1999: Niedersorbisch-deutsches Wörterbuch. Bautzen.
Stefanowitsch, Anatol 2020: Corpus Linguistics. A Guide to the Methodology. Berlin. DOI:10.5281/zenodo.3735822.
Šrejdaŕ, Juro; Zakaŕ, Viktor 2017: Pó serbsku! Gramatika za wuknjecych. Grammatik für Lernende. Bautzen.
Šwjela, Bogumił 1961: Dolnoserbsko-němski słownik. Budyšyn.
Taylor, Charlotte; Marchi, Anna (Hgg.) 2018: Corpus Approaches to Discourse. A Critical Review. London-New York.
Wölke, Sonja 2006: Aktualne tendencje rozwojowe w górnołużyckim języku literackim, in: Zeszyty Łużyckie 39/40, S. 37– 49.
Wornar, Edward 2001: K poměrej mjez hornjoserbskimi nominalnymi kompozitami a jich němskimi předłohami, in: Lětopis 48/1, S. 5–12.

Anlage: Vollständige Liste der 2019 analysierten Texte

Obersorbische Texte		Druckseiten	Textwörter
Os-1	Serbske Nowiny : njewotwisny wječornik za serbski lud[144]	1 073	1 589 919
Os-2	Rozhlad : serbski kulturny časopis[145]	428	160 990
Os-3	A srjedźa Kaponica / wud. Marko Grojlich[146]	400	51 514
Os-4	Doma we wućekach 2. Ze zapiskow, listow a pojednanjow. 1990-2018 / Benedikt Dyrlich	376	89 942
Os-5	Dźiwadło njewidźomnych dźěći / Marcin Szczygielski, pśeł. Jan Měškank	264	59 787
Os-6	Findus a hara z honačom / Sven Nordqvist, pśeł. Diana Šołćina	32	2 612
Os-7	Jank a Majka w njebjesach / Dorothea Šołćina	60	7 832
Os-8	Mišter Krabat. Dušny serbski kuzłar / Měrćin Nowak-Njechorński	48	10 347
Os-9	Nowe dyrdomdeje na wsy / Štefan Paška	52	3 635
Os-10	Paternoster. Teksty młodych awtorow 8	60	10 103
Os-11	Serbska protyka 2020 / red. Pětr Šołta	160	49 028
Os-12	Serbske pomniki. Přewodnik po serbskich wopomnišćach / Trudla Malinkowa	220	107 317
Os-13	W času zeza časa. Popady a wobrady / Róža Domašcyna[147]	116	9 015
Os-14	Warimy z Tomašom / Tomaš Lukaš	134	14 468
Os-15	Wšědne hesła Ochranowskeje bratrowskeje wosady na lěto 2020 / zest. Hinc Šołta	116	19 847
Os-16	Z Lipicy do hole. Wobrazy a powěśće z dawnych časow / zest. Michał Anders, Pětr Lipič	183	34 889
Os-17	Znaki pominaki kopolaki / Měrana Cušcyna, Róža Domašcyna, Měrka Mětowa	56	6 135
gesamt		**3 778**	**2 227 380**

[144] 200 Ausgaben (Nr. 48, 50–248), einschließlich der Wochenendbeilage „Předźenak“ (43 Ausgaben); Festiwalna wosebita přiłoha (1).
[145] Elf Ausgaben (Nr. 1–12), hier ohne Artikel in niedersorbischer Sprache.
[146] Nur obersorbische Seiten der zweisprachigen Ausgabe.
[147] Ohne deutsche Kapitel.

Niedersorbische Texte		Druckseiten	Textwörter
Ns-1	Nowy Casnik : Tyźenik za serbski lud[148]	332	410 211
Ns-2	Rozhlad : serbski kulturny časopis[149]	101	27 068
Ns-3	Bóže słowo na kuždy źeń 2020. Gronka ochranojskeje bratšojskeje wósady / zest. Hartmut S. Leipner	120	20 017
Ns-4	Findus a kokot-spiwarik / Sven Nordqvist, pśeł. Madlena Norberg	32	2 915
Ns-5	Kšet Knut a bur Žur / Jurij Koch	28	2 630
Ns-6	Mejstaŕ Krabat. Dušny serbski guslowaŕ / Měto Nowak-Njechoŕński, pśeł. Wylem Bjero	48	11 149
Ns-7	Serbska pratyja 2020 / red. Horst Adam, Adelheid Dawmowa, Ingrid Hustetowa	176	50 087
Ns-8	Wósłonki mójogo žywjenja / Erwin Hanuš	200	46 662
gesamt		**1 037**	**570 739**

[148] 41 Ausgaben (Nr. 11–13, 15–52).
[149] Elf Ausgaben (Nr. 1–12), hier nur die Artikel in niedersorbischer Sprache.