Marco Scheider

Desiderata der Arbeit mit elektronischen Corpora
Am Beispiel der DWDS-Datenbasis für das 20. Jahrhundert

An der Berliner Akademie sind die lexikographischen Vorhaben in dem vor einigen Jahren gegründeten „Zentrum Sprache“ zusammengefaßt worden; zu diesem Zentrum gehören u.a. die Arbeitsstelle für die DWB-Neubearbeitung, das Goethe-Wörterbuch und das „Digitale Wörterbuch der deutschen Sprache“ (DWDS). Während das Goethe-Wörterbuch noch bis 2025 im Verbund mit den beiden anderen Arbeitsstellen in Hamburg und Tübingen mit der Fertigstellung des Projekts beschäftigt sein wird, zeichnet sich für die DWB-Mitarbeiter/innen ein früherer Umbruch ab: nach der Neubearbeitung der Buchstaben A- F wird die weitere lexikographische Arbeit ab 2013 unter dem Dach des DWDS angesiedelt sein. Das DWDS selbst läuft in seiner ersten von insgesamt drei Phasen à sechs Jahren bereits seit 2007 als Akademienvorhaben.

Neben dem Modul „Aussprache“, für das schon ein fortgeschrittenes Stadium der Bearbeitung vermeldet werden kann, und verschiedenen Schritten zur Vernetzung bereits vorhandener lexikographischer (elektronischer) Ressourcen soll beim DWDS in der ersten Projektphase bis 2012 nicht zuletzt eine ausreichende Datengrundlage für die nachfolgende lexikographische Arbeit geschaffen werden; bis jetzt bietet das DWDS, das schon vor 2007 als Projekt, allerdings noch nicht als Akademienvorhaben an der BBAW angesiedelt und darüber hinaus auf das 20. Jahrhundert konzentriert war,[1] für die Zeit vor 1900 noch kein Corpus. Dieser Lücke soll vor allem mit dem DFG-Projekt „Deutsches Textarchiv“ (DTA) begegnet werden, das im Laufe von sechs Jahren bis zu 1500 Texte aus der Zeit von 1600/1650 bis 1900 digitalisieren und entsprechend aufbereiten wird.

Die DWB-Arbeitsstelle hat im vergangenen Jahr eine Stellungnahme verfaßt, in der Chancen und Möglichkeiten, die in der bevorstehenden Zusammenarbeit mit den Computerlinguisten im Hause liegen, ausdrücklich begrüßt, in der aber auch Desiderata benannt werden, die sich nach den in den letzten Jahren gesammelten Erfahrungen der Arbeit mit dem DWDS-Copus

[1] Daher bis heute ein entsprechendes Logo auf der Website www.dwds.de (23.4.09), das auch auf den beigefügten Screenshots zu erkennen ist.

als nicht unerheblich erweisen. Einige der dort vorgetragenen Argumente sollen hier mit Beispielen veranschaulicht werden. Für eine genauere Erörterung der Perspektiven der historischen Lexikographie nach 2012 wäre die Diskussion naturgemäß eigentlich anhand des DTA zu führen; da dessen Ergebnisse aber noch nicht vorliegen, die Quellen also noch nicht nutzbar sind, wird stellvertretend die DWDS-Datenbasis für das 20. Jahrhundert herangezogen. Die vorgelegten Befunde müssen dabei in Relation zur Corpusgröße gesetzt werden; während das systematisch zusammengestellte, ausgewogene sogenannte „Kerncorpus“ des DWDS, das öffentlich zugänglich ist, 100 Millionen laufende Textwörter beinhaltet und sich als Referenzcorpus für das 20. Jahrhundert versteht, dürfte das auf dem DTA-Projekt basierende Corpus für die früheren Jahrhunderte weniger umfänglich sein. Immerhin wird die Digitalisierung, so ist zu hoffen, dem DTA unter rechtlichen Gesichtspunkten keine größeren Schwierigkeiten bereiten, während im DWDS-Corpus aus urheberrechtlichen Gründen bedeutende Autoren des 20. Jahrhunderts bis jetzt nur unzureichend (z. B. Thomas Bernhard, Peter Handke) oder in hinter heutigen philologischen Standards zurückbleibenden Ausgaben (z. B. Kafka) vertreten sind. Hier sollten Gespräche mit Verlagen baldmöglichst zu den notwendigen Ergebnissen führen.

Um Mißverständnisse auszuschließen: es handelt sich nicht um eine Kritik am DWDS, dessen Ansatz ohne Frage in die Zukunft weist; vielmehr geht es um Hinweise darauf, wo noch Probleme bestehen, nach deren Lösung in nächster Zeit gesucht werden sollte. Diese Hinweise beruhen im übrigen nicht auf einer methodisch-theoretischen Reflexion, sondern auf der lexikographischen Praxis der DWB-Neubearbeitung, weswegen die Beispiele einem schmalen alphabetischen Abschnitt am Ende des *A* bzw. am Anfang des *B* entspringen. Sie beziehen sich unter verschiedenen Gesichtspunkten auf die Grundlagen lexikographischen Arbeitens. Die Phänomene, die dabei in den Blick geraten, werden nur mit Einzelfällen illustriert, die aber jeweils als repräsentativ für eine Reihe gleichgelagerter Befunde gelten können.

Das erste Problem, um das es gehen soll, ist das der automatischen **Lemmatisierung**. Für das 20. Jahrhundert wird es angesichts normierter Schreibungen überwiegend als gelöst angesehen, wobei bestimmte Fehlertypen wie nicht erkannte Präfixbildungen anerkanntermaßen weiterhin eine Herausforderung darstellen. Deutlich schwieriger ist die Lage natürlich für die nicht normierten Texte früherer Jahrhunderte. In den auf Folie 1 gezeigten Artikeln aus der DWB-Neubearbeitung zu *ausschlaufen*, *[1]ausschleifen*,

[2]*ausschleifen* und *ausschliefen* sind die bei der Bearbeitung aufgetretenen Probleme der Zuordnung von Belegen anhand der aufgeführten Nebenformen noch teilweise erkennbar, in aller Deutlichkeit treten sie aber erst in der tabellarischen Übersicht auf Folie 2 zutage: das Durcheinander von Rundung, Entrundung, Diphthongierung und Umlautbildung dürfte einstweilen jedes Lemmatisierungsprogramm in die Knie zwingen, weswegen bei diesem Arbeitsschritt auf absehbare Zeit nur an ein händisch kontrolliertes Verfahren zu denken ist. Natürlich ließe sich einwenden, daß es sich bei den genannten Formen um besonders kniffelige Beispiele handelt; daß aber auch hinter sehr viel geläufigeren Lemmata ähnlich diffizile Probleme lauern können, zeigen die Folien 3-5, auf denen DWB-Belegzettel mit unregelmäßigen Formen zu *ausschreien* abgebildet sind, die sich bis auf weiteres jeder automatischen Lemmatisierung entziehen. Die z.T. schon laufende Arbeit an geeigneten Programmen für die Analyse historischer Wortformen ist also zu intensivieren, und sie sollte künftig am besten in Abstimmung mit den DWB-Mitarbeitern geleistet werden.
Damit verbunden ist ein Aspekt, den mein Kollege Harry Fröhlich mit der eingängigen Formel auf den Punkt gebracht hat: im Archiv *finde* ich, in der Datenbank *suche* ich. Die Frage ist nämlich, wie ich in einer Datenbank an diese abweichenden Formen herankommen könnte. Bei einem Zettelkasten ist es ja so, daß der Bearbeiter irgendwann auch auf die unregelmäßigen Formen als Belegzettel stößt, selbst wenn sie ggf. falsch eingeordnet sein sollten. Solange aber die automatische Lemmatisierung nicht hundertprozentig verläßlich arbeitet, werden mir diese devianten Formen bei der elektronischen Abfrage gar nicht angezeigt, und ich müßte in der Datenbank händisch nach ihnen suchen. Woher aber soll ich überhaupt wissen, wonach ich genau suchen soll?

Neben der Lemmatisierung bildet natürlich auch die korrekte **Datierung** eines Belegs eine elementare Grundlage für seine Bearbeitung. Welche Probleme sich in dieser Hinsicht bei der automatisierten Texterfassung ergeben können, zeigte eine Suche nach einem Letztbeleg zum Stichwort *Avertissement* im DWDS-Corpus (Folien 6-7[2]): mindestens acht der neun angezeigten Treffer, die maschinell zwischen 1908 und 1979 datiert werden, gehören in das 19. oder gar 18. Jahrhundert, es handelt sich jeweils um Zitate; die historisierende Schreibweise in Beleg 6 ist ein Grenzfall. Angesichts solcher Ergebnisse ist also dringend danach zu fragen, wie diese systembedingten Fehler vermieden werden könnten, zumal längst nicht alle in Texte eingestreuten Zitate so eindeutig als solche zu erkennen sind wie im

[2] Alle DWDS-Screenshots von www.dwds.de (23.4.09).

vorgeführtenen Beispiel. Ist auch hier eine händische Kontrolle nach dem Muster der traditionellen Exzerption die einzige Möglichkeit der Fehlerminimierung oder sind dafür andere Lösungen denkbar?

Zur Frage des **Lemmabestands**: für die Erstellung des Artikels zum seit dem 16. Jahrhundert bezeugten, heute überwiegend partizipial begegnenden Verb *auswichsen* bot das DWB-Archiv zwar eine Reihe von Belegen für das 20. Jahrhundert, literarische Belege fanden sich allerdings nur bis zum Jahr 1954, für die Dokumentation bis in die jüngste Zeit wurde letztlich auf Wörterbuchbelege zurückgegriffen (Folie 8). Der naheliegende Versuch, dem Mangel an literarischen Belegen mittels einer Befragung des DWDS-Corpus abzuhelfen, schlug fehl: die einfache Suche *auswichsen* ergab 0 Treffer (Folie 9). Bei der Suche nach dem Partizip *ausgewichst* wurden zwar 117 Treffer angezeigt (Folie 10); sie wiesen aber allesamt die Form *ausgewichen* aus, gehörten also zu *ausweichen*. Die exakte Suche nach der Partizipialform mit *@ausgewichst* blieb wiederum ohne Treffer (Folie 11). Anscheinend ist das Lemma also im DWDS-Corpus nicht vorhanden – wenn man nicht Fehler bei der vielleicht auch für das 20. Jahrhundert nicht immer unproblematischen Lemmatisierung annehmen mag, auf die zumindest das Ergebnis der unscharfen Suche nach dem Partizip hinweist. Der Befund ist nicht nur deswegen erstaunlich, weil den 100 Millionen laufender Textwörter für das 20. Jahrhundert im DWDS-Corpus lediglich drei Millionen Belege aus vielen hundert Jahren im DWB-Archiv für die Buchstabenstrecke *A*-*C* gegenüberstehen, sondern weil DWDS-Statistiken von 2,2 Millionen verschiedenen Textwörtern im Kerncorpus ausgehen[3], was zunächst die Erwartung wecken könnte, jedes im DWB bezeugte Lemma müsse sich auch im DWDS nachweisen lassen, besonders wenn es sich nicht um Hapaxlegomena handelt.

Ein ähnlicher Fall ergab sich bei *Beding*, das im DWB-Archiv im 20. Jahrhundert mit sechs Belegen vertreten ist (Folien 12-17): der letzte DWB-Beleg stammt von 1935. Die Suche nach einem späteren Beleg im DWDS-Corpus brachte keinen Erfolg: 15 der 16 Belege boten *Bedingen*, nicht *Beding* (Folie 18), und auch der einzige scheinbar korrekte Beleg erwies sich als – maschinell schwer zu erkennender – Fehlläufer: die entsprechende Textstelle stammt aus einer Annonce, *Beding* ist hier Abkürzung für *Bedingung*, der Punkt ist entweder bei der Digitalisierung entfallen oder war schon im Druck nicht vorhanden (Folie 19). Heißt das nun,

[3] Vgl. A. Geyken: Korpora als Korrektiv für einsprachige Wörterbücher. In: LiLi 136 (2004), S. 72-100, hier S. 88. Im selben Heft findet sich der Aufsatz von Wolfgang Klein, den Thomas Gloning zuvor referiert hat.

daß sich *Beding* im weiteren Verlauf des 20. Jahrhunderts nicht mehr nachweisen läßt? Als weitere Anlaufstelle für diese Frage diente zunächst das am IdS angesiedelte „Online-Wortschatz-Informationssystem Deutsch“ (OWID), das ja ebenfalls auf die jüngere Sprache konzentriert ist. Dort aber fanden sich nur Hinweise auf Schreibung und Worttrennung des Lemmas sowie ein Link auf canoo.net (Folie 20), und dieser Link führte nun zum Verb *bedingen*, keineswegs zum Substantiv (Folie 21). Sofern dieses von OWID praktizierte Verfahren, zunächst Basisinformationen und Verknüpfungen zu anderen Informationssystemen anzubieten, als eine beispielhafte Umsetzung der von W. Klein in seinem Aufsatz propagierten inkrementellen Funktionalität digitaler lexikalischer Systeme verstanden werden kann, ist ihm wohl einstweilen mit Skepsis zu begegnen. Letztlich aber kam doch aus Mannheim, bei der Befragung von COSMAS, u.a. mit einem Beleg aus der „Neuen Kronen-Zeitung“, der Nachweis jüngerer Verwendung des im süddeutschen / schweizerischen / österreichischen Sprachraum wohl noch geläufigen Wortes; auch COSMAS bot freilich einige Fehlläufer wie die aus dem Anzeigenteil des „Neuen Deutschland“ (Folie 22).

Bei einem Kompositum stehen bekanntlich häufig verschiedene Bildungstypen entweder gleichberechtigt oder mit abgestufter Frequenz nebeneinander, wobei in der Regel zuletzt eine Form dominiert. Bei dem in dieser Hinsicht symptomatischen Artikel *Ausnahmeerscheinung* war zu prüfen, ob es für die **Nebenform** *Ausnahmserscheinung* im DWDS-Corpus noch einen späteren Beleg als im DWB-Archiv gibt, also (Folie 23) nach 1927. Dies war nicht der Fall: es fand sich vielmehr überhaupt kein Beleg für diese Form (Folie 24).

Auch in puncto **Phraseologie** sind immer wieder überraschende Befunde zu konstatieren. In dem DWB-Material zum Artikel *Bauernjunge*, das an sich keine spektakulären Erkenntnisse bereithielt, fiel zumindest die Verbindung *es regnet* bzw. *weht Bauernjungen* auf, die für das 20. Jahrhundert dreimal bezeugt ist (Folie 25 zeigt einen Entwurf des noch nicht abgeschlossenen Artikels mit Korrekturen). Naturgemäß wurde auch dazu das DWDS-Corpus befragt. Daß es keine Treffer gab (Folie 26), erstaunt auch deswegen, weil die DWB-Belege von Lily Braun und Klabund ja keineswegs als exotisch gelten können. Nach Hinweisen von verschiedenen Seiten, daß die Variante *es regnet Schusterjungen* deutlich häufiger und heute noch dem einen oder anderen Sprachteilnehmer vertraut sei, wurde probehalber dazu eine Abfrage gestartet, die aber ebenfalls keine Treffer erbrachte; es zeigte sich bei der alleinigen Suche nach dem Substantiv sogar (Folie 27), daß es für den *Schusterjungen* im Corpus

überhaupt nur ein gutes Dutzend Nachweise gibt; davon stammt der letzte angezeigte aus dem Jahr 1939, keiner weist den Gebrauch des Wortes in der Druckersprache nach, kein einziger kennt das Brötchen.

Im Bewußtsein der häufig nur relativen Gültigkeit, aber dennoch mit Engagement wird beim DWB die Suche nach **Erstbelegen** für Lemmata als solche oder für spezielle Bedeutungen betrieben. Soweit es sich um Lemmata des 20. Jahrhunderts handelt, sollte das DWDS für die entsprechende Dokumentation eigentlich prädestiniert sein. Die Befunde zeugen aber nicht von der erwarteten Verläßlichkeit. So hat die Suche nach einem Erstbeleg für den im DWB-Archiv seit 1927 bezeugten *Baggersee* (Folie 28) gezeigt, daß das DWDS das Lemma erst 70 Jahre später nachweist (Folie 29). Beim – im Entwurfsstadium befindlichen – Artikel *Bauernopfer* ist der entsprechende DWB-Erstbeleg nicht nur über 60 Jahre älter als der des DWDS, in letzterem ist auch die Herkunft aus dem Schachspiel gar nicht nachweisbar, das ja erst die Grundlage für die heute verbreitete Übertragung bildet (Folien 30/31).

Als letzter Punkt schließlich soll hier in Analogie zu den Arbeitsschritten bei der Erstellung eines DWB-Artikels die Frage der **Belegüberprüfung** angesprochen werden. Bei der Exzerption wie bei der elektronischen Texterfassung schleichen sich immer Fehler ein, in dieser Hinsicht dürfte es keinen Unterschied zwischen händischen und maschinellen Verfahren geben. Daher ist es unerläßlich, die Belege noch einmal auf ihre Korrektheit zu überprüfen. Die Dringlichkeit dieses Arbeitsschritts läßt sich anhand der Artikelarbeit zur Lehnübersetzung *Außenseiter* demonstrieren. Der Beleg aus Borchardts „Jettchen Gebert“ wird im DWDS zitiert als *Wir sind die Außenseiter im Derby*. Tatsächlich aber ergab die Überprüfung der Quelle in der DWB-Bibliothek, daß die entsprechende Stelle lautet *Wir sind die Außenseiter von Derby* (Folie 32), was für die Herleitung eine entscheidende Differenz bedeutet. Zu fragen ist also, wie eine adäquate Überprüfung der Belege aus digitalen Corpora zu bewerkstelligen ist. Sollten die Bearbeiter/innen immer ein Image zur Verfügung und im Blick haben oder sind noch andere Methoden der Qualitätssicherung, vielleicht schon in einem früheren Stadium der Texterfassung, denkbar?

Um es noch einmal zu sagen: es ging hier nicht darum, dem DWDS Versäumnisse anzukreiden. Ziel aller Überlegungen muß vielmehr die Qualitätssicherung der zukünftigen Arbeit sein. Im Moment zeigen die Befunde einfach, daß noch einiges an Vorbereitung zu

leisten ist, soll der von den Wörterbuchvorhaben der BBAW bzw. im gesamten Akademienprogramm gesetzte philologische Standard auch bei der lexikographischen Arbeit im DWDS erreicht werden – wobei die aufgezeigten Phänomene Auswirkungen sicher auch auf viele andere Nutzungsarten und Untersuchungen hätten. Wie kann nun eine Qualitätssicherung gewährleistet werden?

Im DWDS selbst ist schon vor einiger Zeit für die hausinterne Nutzung rund um das Kerncorpus ein erweitertes, opportunistisches Corpus aufgebaut worden, mit dem die Zahl der laufenden Textwörter längst die Milliardengrenze überschrittten hat. Allerdings speist sich diese Erweiterung überwiegend aus Zeitungen der achtziger und neunziger Jahre[4], deren begrenzte Aussagekraft sich schon daran zeigt, daß sie die Ergebnisse für die oben präsentierten Beispiele überwiegend nicht oder nur unwesentlich verbessert hätte. Außerdem, und das ist für die historische Lexikographie maßgeblich, läßt sich eine Erweiterung der Textbasis für die früheren Jahrhunderte eben nicht in vergleichbarere Weise durch den Ankauf von Jahrgangs-CDs realisieren. Dem DTA muß also entweder die Möglichkeit gegeben werden, über die 1500 Titel hinaus weitere Texte zu erfassen, oder die Textgrundlage muß über Kooperationen erweitert werden. Mit der jetzigen Grundlage, die im Umfang hinter dem des DWDS-Kerncorpus zurückbleibt, ist historische Lexikographie nicht zuverlässig zu betreiben.

Das DWB hat naturgemäß seit jeher andere Maßnahmen ergriffen. Wer das DWB-Archiv kennt, weiß, daß es keineswegs frei von Lücken und Unzulänglichkeiten ist, dafür gibt es bei der Artikelarbeit aber auch festgelegte Abläufe zum Abgleich des Materials mit anderen wichtigen Wörterbüchern und Nachschlagewerken, mit denen ggf. das Fehlen von Belegen wie etwa beim oben genannten *auswichsen* kompensiert wird. Sind vergleichbare Verfahren für das DWDS denkbar? Wird es dann alle nötigen Informationen in digitaler Form geben oder wird man auf eine Mischung von papierenen und digitalen Ressourcen zurückgreifen? Diese Fragen sollten in der Runde diskutiert werden.

[4] Vgl. dazu wiederum A. Geyken in dem oben erwähnten Aufsatz S. 82.