Ein Korpus als Garant zuverlässiger lexikografischer Informationen? Eine vergleichende Stichprobenuntersuchung

Ulrich Schnörch & Petra Storjohann

Schlüsselwörter: *Zuverlässigkeit, Authentizität, Korpusmethoden, Analysemethoden, Arbeitsgrundlagen*.

Abstract

Current working practice of established German dictionaries incorporates large corpora as the basis of most analyses, descriptions and presentations. It is, however, individual lexicological and/or different corpus-methodological approaches that play a crucial role in the process of extracting and documenting lexicographic information in individual reference works. This paper addresses the question of how reliable information is in some electronic German dictionaries. Objects of our investigation are different types of corpus dictionaries, e.g. a digitized dictionary, a reference work that compiles its data fully automatically, a lexicographic system combining different electronic resources, and a corpus-assisted dictionary that examines and interprets its corpus data lexicographically. Critical examinations of such reference works inevitably come up with questions of authenticity and reliability of the given dictionary information. The advantages and disadvantages of various lexicographic or corpus-linguistic methods which are individually implemented will be outlined and critically analyzed with the help of examples. According to an extensive study (cf. Müller-Spitzer 2011) reliability of given information is one of the key criteria assigned to any reference work by users. We will elicit how different corpus methods expose different descriptions of natural discourse and how they answer questions of authenticity, typicality and reliability with regard to phenomena such as meaning spectrum, collocations, antonymy and hyperonymy. Overall, this paper is a critical account of the current German lexicographic developments. It will include discussions on meta-lexicographic demands and focus on whether there are suitable complementary corpus approaches providing authentic dictionary information to a satisfactory extent.

1. Einleitung

In diesem Vortrag wird die Zuverlässigkeit von Informationen in verschiedenen deutschen elektronischen Wörterbüchern stichprobenartig untersucht. Den Wörterbüchern ist gemeinsam, dass die jeweiligen Angaben auf der Auswertung von Korpusdaten basieren, aber unterschiedliche lexikologische Ansätze verfolgen und diverse computerlinguistische Methoden und Verfahren heranziehen. Im Fokus der Betrachtung stehen digitalisierte Aufbereitungen von Printwörterbüchern (z. B. Duden-GWDS bzw. Duden online), automatisiert zusammengestellte Wörterbücher (z. B. Wortschatzprojekt Leipzig), digitale lexikalische Systeme, die unterschiedliche Ressourcen elektronisch verbinden (z. B. DWDS) und Nachschlagewerke, die Korpusdaten lexikografisch auswerten (z. B. *elexiko*). Anhand unterschiedlicher Informationstypen, wie Bedeutungsspektrum, Gegensatzwörter und Hyponyme zeigen wir, wie die verschiedenen Projekte diese Angaben gewinnen und dokumentieren, und welche Defizite dabei beobachtet werden können. Dies führt schließlich zu der grundlegenden Frage nach der Authentizität der Wörterbücher, die prinzipiell auf Korpusbasis erstellt werden.

2. Authentizität von Angaben

Die Hauptfunktion eines Wörterbuches ist es ganz allgemein, zuverlässige Informationen zu bestimmten Fragestellungen bereitzustellen. Zunächst ist also zu fragen, was ein zuverlässiges Wörterbuch ausmacht. Inwieweit wird die Realität der deutschen Wörterbuchlandschaft den metalexikografischen Ansprüchen bzw. NutzerInnenwünschen gerecht. Als Ausgangspukt unserer Untersuchung dient folgende Bemerkung von Atkins/Rundell (2008: 45):

> A reliable dictionary is one whose generalisations about word behaviour approximate closely to the ways in which people normally use language when engaging in real communicative acts.

Sprecher führen die Authentizität lexikografischer Angaben als eines der wichtigsten Kriterien bei der Qualitätsbeurteilung eines Wörterbuchs an. Das bestätigt auch eine am Institut für Deutsche Sprache durchgeführte Benutzerstudie aus dem Jahr 2010 (siehe Müller-Spitzer u.a. 2011). Es besteht heute außerdem Konsens darüber, dass authentische Informationen zu sprachlichen Phänomenen tatsächlichen Sprachgebrauch im kontextuellen Diskurs illustrieren, wie er anhand von umfangreichen sprachlichen Daten, sprich einem Korpus, musterhaft erkennbar wird. Den Anspruch an verlässliche Beschreibungen machen dabei gängige Wörterbuchprojekte oft in ihren Umtexten explizit, wie etwa Duden-GWDS:

> ‚Das große Wörterbuch der deutschen Sprache' […] ist die umfassende und authentische Dokumentation der deutschen Sprache vor dem Übergang ins neue Jahrtausend. (Vorwort, Bd. 1, S. 5)

Die Beschreibung natürlich vorkommender Sprache ist generell zentrales Ziel sprachlicher Untersuchungen seit der Verfügbarkeit großer elektronischer Korpora. Korpora ermöglichen es, zahlreiche auch unerwartete Beobachtungen machen zu können. Hanks (1990: 40) charakterisiert eine Wesensart der Sprache wie folgt: "Natural languages are full of unpredictable facts […] which a corpus may help us to tease out". Anhand von Stichproben soll nun gezeigt werden, dass zuverlässige und authentische Informationen nicht ausschließlich mit Korpusdaten lexikografisch benannt werden können. Die Arbeitsmethoden in den gängigen deutschen Wörterbüchern weichen erheblich voneinander ab, was auch Konsequenzen für die Zuverlässigkeit der bereitgestellten Informationen hat.

Prinzipiell ist es die Herausforderung der Lexikografie, bei der Arbeit mit einem Korpus aus all der Vielfalt sprachlicher Realisierungen in Massendaten das Typische, Musterhafte und das generell Verbreitete an Strukturen herauszufiltern. Geleitet von dieser Hypothese wurden unterschiedliche Einträge aus mehreren gegenwartssprachlichen monolingualen deutschen Wörterbüchern mit Korpusdaten verglichen. Als Vergleichsinstrument diente uns die weltweit größte Sammlung elektronischer Korpora der deutschen Gegenwartssprache das so genannte Deutsche Referenzkorpus (DeReKo) am Institut für Deutsche Sprache in Mannheim mit über 5 Milliarden Textwörtern.

3. Beispiele

3.1. *Angaben zum Bedeutungsspektrum*

Ausdrücke mit soziopolitischer Brisanz werden in gegenwartsprachlichen Korpora sehr vielfältig diskutiert. Diese oftmals hochfrequenten Wörter, wie etwa *Mobilität*, sind Diskursschlüsselwörter mit sehr facettenreichen Verwendungsmustern und semantischen Fokussierungen. Das Stichwort **Mobilität** wird im Duden-DWDS in Bezug auf sein Bedeutungsspektrum in 3 Einzelbedeutungen untergliedert und wie folgt beschrieben:

Mo|bi|li|tät, die; - [lat. mobilitas, zu: mobilis, mobil]: **1.** (bildungsspr.) *[geistige] Beweglichkeit:* Die M. der Vierziger ist drastisch eingeschränkt (Schreiber, Krise 37); seine Argumentationen zeugten von hoher M. **2.** (Soziol.) *Beweglichkeit (in Bezug auf den Beruf, die soziale Stellung, den Wohnsitz):* die soziale, regionale M. der Arbeitnehmer; Bei der Arbeitssuche werden mehr M., geringere Lohn- und Gehaltsforderungen ... empfohlen (Saarbr. Zeitung 2.10.79, 4); eine Gesellschaft mit hoher M. **3.** (Milit. selten) *mobiler Zustand, Kriegsbereitschaft:* eine Demonstration der hohen M. ... der sowjetischen Kriegsmarine (Bundestag 190, 1968, 10325). (Quelle: Duden-GWDS auf CD-ROM)

Für die Untersuchung der zugrunde liegenden Korpusdaten wurde das Recherchesystem COSMAS II mit der darin befindlichen Software ‚Statistische Kollokationsanalyse und –clustering' genutzt. Die Analyse von unmittelbaren Kontextwörtern, so genannten Kollokatoren bzw. Kookkurrenzen verdeutlicht, wie eingeschränkt oder sogar lückenhaft die lexikografische Darstellung der Bedeutung ist. Das Resultat der Analyse statistisch signifikanter Gebrauchsmuster zeigt, dass die dort erfassten beiden ersten Einzellesarten ‚[geistige] Beweglichkeit' und ‚(soziol.) Beweglichkeit (in Bezug auf den Beruf, die soziale Stellung, den Wohnsitz)' nicht die zentralen Verwendungsaspekte reflektieren. Insbesondere Substantiv- und Adjektivkollokatoren und dazugehörige komplexere syntagmatische Konstruktionen fungieren als Indikatoren für ein breiteres semantisches Spektrum (siehe Tabelle 1).

Tabelle 1. Auswahl an signifikanten Kollokatoren als Indikatoren für weitere Lesarten von *Mobilität*.

Kollokatoren	**Syntagmatische Muster**
eingeschränkter	älteren Menschen ... eingeschränkter Mobilität
umweltfreundliche	umweltfreundliche und nachhaltige Mobilität im Stadtteil
nachhaltige	nachhaltige Mobilität und Klimaschutz
Kommunikation	Mobilität ... Kommunikation
mangelnde	mangelnde Mobilität im ländlichen
Verkehrsteilnehmer	Mobilität aller Verkehrsteilnehmer
Ballungsräumen	Mobilität in Ballungsräumen
Erreichbarkeit	Mobilität [und] Erreichbarkeit
Verkehrssicherheit	der Mobilität und der Verkehrssicherheit
motorisierte	die motorisierte Mobilität
Radwege	Mobilität und Radwege
körperliche	körperliche Mobilität
barrierefreie	für barrierefreie Mobilität
Infrastruktur	Mobilität [und] Infrastruktur
Rädern	für der|die Mobilität auf vier|zwei Rädern
Vernetzung	Mobilität [und ...] Vernetzung
Patient	eingschränkte Mobilität der Patienten
Körperpflege	Bereichen Körperpflege, Mobilität und Ernährung

Aspekte wie ‚körperliche Flexibilität' (z. B. signalisiert durch *eingeschränkt, körperlich, Patient, Körperpflege*), ‚flexible Erreichbarkeit für Kommunikation' (z. B. signalisiert durch *Kommunikation, Erreichbarkeit, Vernetzung*) sowie ‚Beweglichkeit im Sinne von Transport' (z. B. signalisiert durch *umweltfreundlich, Verkehrsteilnehmer, motorisiert, Verkehrssicherheit, barrierefrei, Infrastruktur*) sind keine peripheren kontextuellen Verwendungen, sondern stellen signifikante Instanzen aus einer Vielzahl verschiedener Textquellen dar und können bzw. müssen zur stabilen Kernbedeutung des Ausdrucks *Mobilität* gerechnet werden. Darüber hinaus sind diese Lesarten frequenter und signifikanter

als die im Wörterbucheintrag aufgeführte Verwendungsweise ‚geistige Beweglichkeit', die aufgrund der Datenlage als die am geringsten belegte Lesart zu bewerten ist. Es handelt sich auch nicht um Verwendungen des Ausdrucks, die sich erst nach Erscheinen des Printwörterbuchs Duden-GWDS von 1999 (der Basis der digitalen Version von 2000) entwickelt haben, sondern die bereits in den Texten der 90-er Jahre deutlich präsent waren. Insbesondere Lesarten wie ‚körperliche Flexibilität', und ‚Beweglichkeit im Sinne von Transport' hätten auch zu diesem Zeitpunkt schon erfasst werden können.

Die 3. Lesart des Dudeneintrags ‚*mobiler Zustand, Kriegsbereitschaft*' wurde mit der Onlineschaltung des Dudens gestrichen (vgl. dazu www.duden-de). Das deutet auf die zwischenzeitliche Aktualisierung des Artikels hin. Der folgende Vergleich zwischen dem Dudeneintrag und der Korpusauswertung, deren Resultate das Wörterbuch *elexiko* dokumentiert, zeigt eine unterschiedliche Bewertung auch hinsichtlich der Anordnung der Lesarten:

Tabelle 2. Vergleich Duden-GWDS mit Ergebnissen der Korpusuntersuchung in *elexiko*.

Duden-GWDS	**Korpus** (Resultat in elexiko)
‚geistige Beweglichkeit'	‚Motorisiertheit'
‚Beweglichkeit (in Bezug auf den Beruf, die soziale Stellung, den Wohnsitz)'	‚berufliche Flexibilität'
	‚Erreichbarkeit'
	‚körperliche Beweglichkeit'
	‚geistige Beweglichkeit'

Auch wenn es, je nach redaktioneller Richtlinie, unterschiedliche Kriterien zur Disambiguierung von Lesarten und unterschiedliche Interpretationen von Kontexten gibt, die stärker getrennt oder zusammengefasst werden können, so handelt es sich aber in diesem Fall nicht um mögliche variable kontextuelle Zuordnungen. Lexikografische Kompetenz und der Einsatz elektronischer Ressourcen spielte bei der Erarbeitung des Nachschlagewerkes eine Rolle. Dieses Beispiel legt jedoch den Schluss nahe, dass zur Erarbeitung des Wörterbucheintrages das jeweilige Korpus nicht korpusgesteuert (corpus-driven) zur Erfassung des Bedeutungsspektrums genutzt wurde. Das heißt, empirische Kontextanalysen von Massendaten wurden vermutlich lediglich zur Verifizierung bereits bestehender linguistischer Annahmen durchgeführt. Die mutmaßlich introspektiv gewonnenen Informationen, in diesem Fall weniger Lesarten, wurden korpusbasiert (corpus-based) mit Belegen aus dem Korpus nachträglich bereichert und „verziert" (vgl. Tognini-Bonelli 2001). Bei einem solchen kompetenzgestützten Vorgehen dient das Korpus als Validierungsinstrument bestehender Hypothesen und als elektronische Belegsammlung zur beispielhaften Illustrierung der lexikografischen Dokumentation, nicht aber als Datengrundlage zur explorativen linguistischen Auswertung. Mit ausschließlich korpusbasierten Verfahren an ein Korpus heranzutreten hat den großen Nachteil, signifikante Muster möglicherweise nicht aufzudecken und ggf. auch aufgrund mangelnder statistischer Auswertungen zu Fehleinschätzungen über die Typikalität oder Usualität eines sprachlichen Phänomens zu kommen. Sprache im Sinne von Hanks (s.o.) lässt sich so nicht zufriedenstellend erfassen und beschreiben.

3.2. *Bedeutung und Synonyme*

Das Digitale Wörterbuch der Deutschen Sprache (DWDS) geht methodologisch einen anderen Weg. Es stellt ein lexikalisches System dar, das verschiedene Angabebereiche aus

vier unterschiedlichen Ressourcen für einen Eintrag zusammenstellt. Der Vorteil ist, dass viele und verschiedenartige Informationen zusammen in verschiedenen Frames (so genannten Panels) bereitgestellt werden (siehe Abbildung 1). Der Nachteil dieser heterogenen Struktur liegt auf der Hand: Der Bezug der einzelnen Angaben zueinander ist nicht immer gewährleistet, weil sich jede darin aufgenommene Ressource ursprünglich wiederum unterschiedlicher Arbeitsgrundlagen und Methoden bediente. Die jeweiligen Daten-Quellen sind selbstredend nur bedingt aufeinander abgestimmt, was deren Vergleichbarkeit bis zu einem gewissen Grad eingeschränkt.

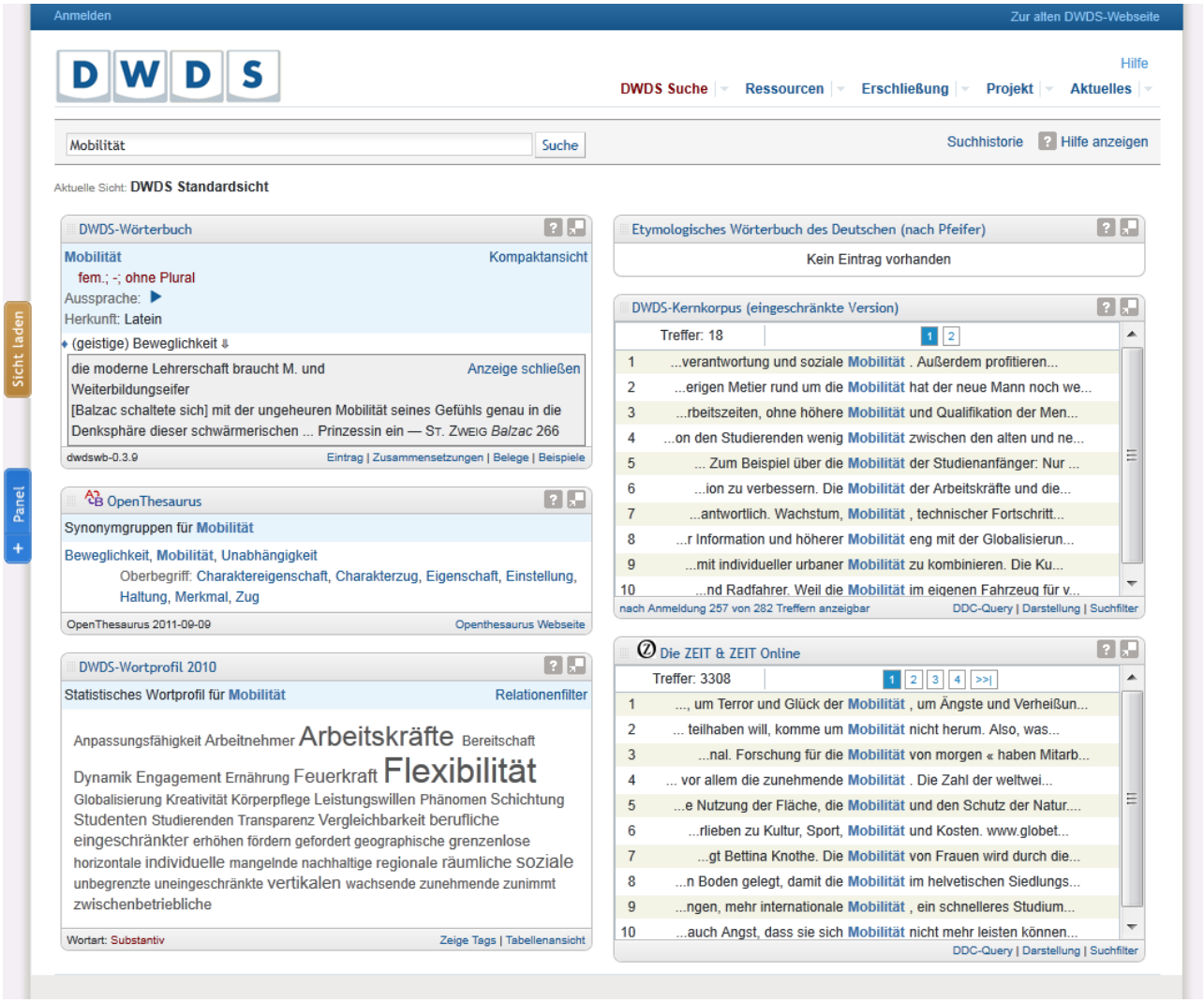

Abbildung 1. Eintrag *Mobilität* im DWDS.

Beim Nachschlagen der Bedeutung des Stichwortes **Mobilität** erhält man die Information, dass es im Sinne der ‚(geistigen) Flexibilität' verwendet wird. Diese Information stammt aus der retrodigitalisierten Fassung des WDG von 1961-1977, das nicht auf der Basis eines elektronischen Korpus erstellt wurde. Darunter befinden sich die Synonymangaben *Beweglichkeit* und *Unabhängigkeit* aus einem separaten Thesaurus, einer Open-Access-Ressource. Dieser Thesaurus ist interaktiv und „jeder kann bei Open Thesaurus mitmachen und Fehler korrigieren oder neue Synonyme einfügen" (Zitat von Webseite: http://www.openthesaurus.de/). Zusätzlich erhalten Nachschlagende ggf. etymologische Angaben aus einem historischen Wörterbuch[1]. Beispiele aus konkreten sprachlichen Kontexten sowie ein lexikalisch-semantisches Wortprofil ergänzen die bereits erwähnten

Informationen. Diese sind die eigentlichen korpusgestützten Wörterbuchangaben. Sie spiegeln aber die Informationen der semantischen Paraphrase inhaltlich nicht wider, sondern stehen ihnen vielmehr als eigenständige Informationsergänzungen zur Seite gegenüber. Für diesen Teil der Angaben liegt ein großes Korpus als Grundlage der lexikografischen Informationen vor. Ausschließlich auf dieser Basis können jedoch zuverlässige und inhaltlich konsistent vernetzte Einträge nicht gewährleistet werden, denn die Korpusanalyse ist per se keine integrative Maßnahme, wenn die Befunde methodologisch nicht rückgekoppelt werden bzw. nicht rückkoppelbar sind, da sie einen anderen Ursprung haben.

Das DWDS muss sich die Frage stellen, wie NutzerInnen die bereitgestellten Informationen aufeinander beziehen bzw. interpretieren. Gegenwartsprachliche Beschreibungen haben bei einem zugrunde gelegten Wörterbuch wie dem WDG für gegenwärtig zentrale Schlüsselwörter etwa *Mobilität* ihre Grenzen. Die Informationen aus dem Kernkorpus dagegen sind aktuell und stellen nicht die Basis des retrodigitalisierten WDG dar. Synonymangaben wiederum kommen aus einer Open-Source-Quelle, die erneut eine andere Datengrundlage als das Korpus hat, diese aber nicht explizit aufführt. Diese Zusammenhänge werden bei näherer Betrachtung der Einträge inhaltlich aufgrund zahlreicher Inkonsistenzen und fehlender Bezugnahmen deutlich; die NutzerInnen werden darauf jedoch eingangs nicht detailliert aufmerksam gemacht.

Das Ziel des DWDS könnte sein, verschiedene Bedürfnisse durch eine gemeinsame, multifunktionale Ressource abzudecken (vgl. Heid 2010: 567), um verschiedene, je nach Bedarf und Nachschlagesituation spezifische Teilwörterbücher anbieten zu können. Die Zuverlässigkeit der gebotenen Informationen kann aber in solchen Fällen nur isoliert für jeden einzelnen Angabebereich in Hinblick auf die dafür genutzte Materialgrundlage beurteilt werden, nicht aber für einen gesamten Artikel im Nachschlagewerk. Unter diesen Vorzeichen kann das lexikalische System eine gute Informationsquelle darstellen, vorausgesetzt, NutzerInnen wissen stets, woher die einzelnen Quellen und Wörterbuchangaben stammen, wie sie zu deuten und miteinander in Zusammenhang zu betrachten sind. Tatsächlich wird NutzerInnen allerdings der Eindruck suggeriert, umfangreiche und vielfältige Informationen über ein Wort gebündelt und kohärent in einem Nachschlagewerk zu erhalten.

3.3. *Antonym- und Hyponymangaben*

Das Wortschatz-Lexikon (Universität Leipzig) ist „Das Nachschlagewerk für Wörter und ihren Gebrauch“, dessen Informationen zu einem Stichwort ausschließlich automatisch extrahiert sind und nicht einer lexikografischen Interpretation der Daten unterliegen. Viele Stichwörter bestechen mit umfangreichen Angaben zu sinnverwandten Ausdrücken, insbesondere zu Antonymen und Hyponymen, die auf den ersten Blick sinnvolle Paare mit korrekten Auflistungen in beiden Richtungen sind, wie etwa *mobil* ↔ *immobil*, *legal* ↔ *illegal*. Bei genauerer Analyse werden aber Inkonsistenzen und sogar Fehler der automatischen Ermittlung deutlich. Softwaregestützte Gegensatzpaaranalysen werden mittels Suche nach Negationspräfixen durchgeführt. Das für das Deutsche produktivste Negationspräfix *un-* bleibt dabei unberücksichtigt. So sind typische Gegensatzwörter, wie etwa *unangemessen* zu *angemessen* oder *unkritisch* zum Eintrag **kritisch** nicht zu finden. Besonders für NichtmuttersprachlerInnen problematisch zu sehen sind Angaben zu Gegensatzpaaren wie *aktiv - passiver*, bei denen ein Ausdruck nicht in seiner Grundform angegeben wird. Bei Substantiven basiert die Suche nach Gegensatzwörtern auf der Analyse des Vorhandenseins der Präfixe *Nicht-*, *Anti-*, und *Wider*. Dieses Verfahren ermittelt aber auch zahlreiche Ausdrücke, die nicht dem usuellen Sprachgebrauch zuzuschreiben sind.

Anhand der Stichprobe **Leben** (siehe Abbildung 2) lässt sich beispielhaft zeigen, dass typische Gegensatzwörter, wie *Ableben*[2], *Absterben*, *Ende*, *Sterben*, *Tod*[3], *Verscheiden* nicht dokumentiert werden. Diese könnten u.a. durch korpusgesteuerte Analysen von Ausdrücken mit ähnlichen Kollokationsprofilen als potenzielle Antonyme ermittelt und anschließend korpusbasiert gezielt überprüft werden. Stattdessen stoßen Nachschlagende auf ungebräuchliche Ausdrücke wie *Nichtleben*, weil auf die lexikografische Interpretation der Daten verzichtet wird.

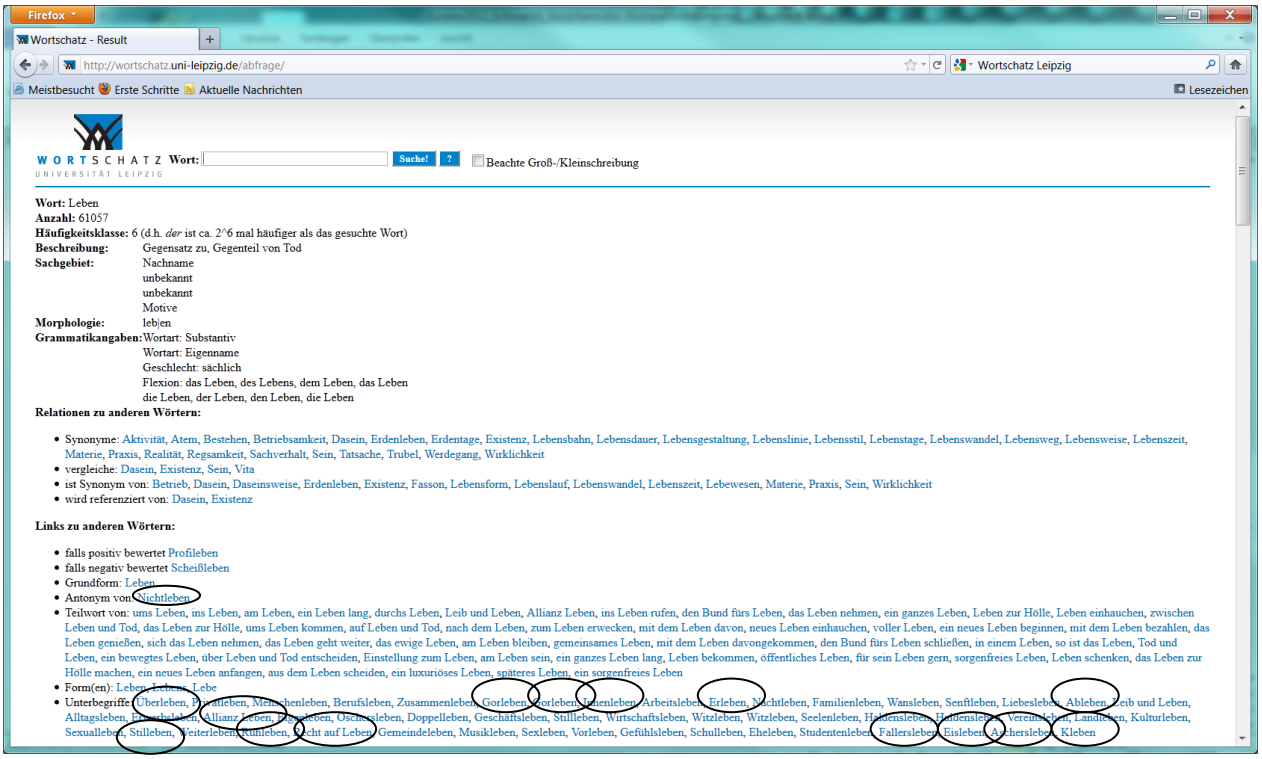

Abbildung 2. Artikel Leben im Wortschatzlexikon mit eigenen Hervorhebungen.

Fehlerhaft sind zahlreiche Bezeichnungen für Hyponyme, die im Beispielartikel *Leben* Städte-, Orts- und Personennamen einschließen z. B. *Aschersleben*, *Eisleben*, *Fallersleben*, *Gorleben*, Produkt- bzw. Firmennamen wie *Allianz Leben*, Ausdrücke wie *Stillleben*, Mehrwortausdrücke wie *Recht auf Leben*, die Derivation *Ableben*, die kein Hyponym ist, sondern bei den Antonymen hätte auftauchen sollen oder Nominalisierungen wie *Erleben* und *Überleben*, die ebenso keine speziellen Arten von *Leben* bezeichnen. Geradezu bizarr ist eine Hyponymangabe wie *Kleben*. Hier werden die automatisierten Analysen nach Zeichenketten zur lexikografischen Sinnlosigkeit.

Das Fehlen jeglicher redaktioneller Auswertung der extrahierten Informationen führt zu zweifelhaften Einträgen und Informationen, die weder am Sprachgebrauch ausgerichtet, geschweige denn authentisch sind. Hier sind introspektiv erarbeitete Synonym- und Antonymwörterbücher jedem korpusbasierten Wörterbuch vorzuziehen, das seine Angaben ausschließlich automatisiert ermittelt. Auch wenn in diesem Fall ein Korpus als Arbeitsbasis eine Fülle an Informationen bereitstellen kann, ist das Korpus nicht im Sinne der Extraktion zuverlässiger Wörterbuchangaben genutzt worden. Das Ziel zuverlässige Angaben zu sinnverwandten Ausdrücken bereitzustellen, wird zugunsten schnell umsetzbarer quantitativer Kriterien aufgegeben.

4. Methodologien und Arbeitsgrundlagen

Susan Hunston (2002: 96) meint „Corpora have so revolutionised the writing of dictionaries“. Das scheint für die englische Lexikografie zu stimmen, bei der Korpora als Quelle für Sprachuntersuchungen und -interpretationen verschiedenster Art genutzt wurden. Der Einsatz umfangreicher Korpora und diverser computerlinguistischer Verfahren hat die Wörterbuchlandschaft inhaltlich verändert, objektivere Sprachbeschreibungen für unterschiedliche Zielgruppen zur Verfügung gestellt. Für die deutsche Wörterbuchlandschaft trifft diese Standortbestimmung nur begrenzt zu. Empirische Analysen werden kaum linguistisch verankert, metalexikografische Forderungen kaum eingelöst.

Zwei Tendenzen werden deutlich. Korpora werden teilweise wenig explorativ genutzt und traditionelle Arbeitsmethoden bevorzugt. In solchen Fällen dient ein Korpus allein als Hilfsmittel um Frequenzen zu ermitteln, um es im Falle lexikografischer Unsicherheit zu konsultieren, um Belege für bestehende Annahmen zu erhalten und um ein bekanntes sprachliches Phänomen quantitativ zu erfassen bzw. zu bewerten. Diese Korpusarbeit ist hinsichtlich bestimmter linguistischer Fragestellungen unverzichtbar. Aber es ist vor allem korpusgesteuerten Analysemethoden, wie dem Einsatz von Kontext- bzw. Kollokationsanalysen zu verdanken, wenn man zu neuen unerwarteten Erkenntnissen über bestimmte Phänomene gelangt, die mit traditionelleren Herangehensweisen nicht erfasst werden (vgl. Sinclair 1997, Tognini-Bonelli 2001, Storjohann 2005).

> There might be a large number of potentially meaningful patterns that escape the attention of the traditional linguist; these will not be recorded in traditional reference works and may not even be recognised until they are forced upon the corpus analyst by the sheer visual presence of the emerging patterns in a concordance page. (Tognini-Bonelli 2001: 86)

Weitgehend Konsens besteht darüber, dass die Erarbeitung eines Wörterbuches heutzutage ein umfangreiches Korpus benötigt. Authentische Beispiele natürlicher Sprache, wie sie im Korpus vorhanden sind, zeugen immer wieder von unvorhersehbaren Fakten über sprachliche Möglichkeiten. Ein Korpus kann dabei helfen, diese ans Licht zu bringen (vgl. Hanks 1990: 40). Weniger ob, sondern vor allem wie LexikografInnen methodologisch mit gewonnenen Korpusdaten umgehen, scheint jedoch bedauerlicherweise zunehmend weniger eine Rolle zu spielen. Eine vergleichbare Entwicklung in puncto abnehmender Methodenreflexion lässt sich auch bei der zweiten Tendenz feststellen: Hier rücken informationstechnologische/softwaregesteuerte Verfahren in den Vordergrund sprachlicher Analysen, weil mit ihnen prinzipiell schnell Massen von sprachlichen Daten hinsichtlich bestimmter Abfragen ermittelt werden können. Dabei spielen zwar korpusgesteuerte Verfahren die entscheidende, aber leider auch ausschließliche Rolle. Schnelle, umfangreiche automatisiert gewonnene Informationen korpusgesteuerter Analysen werden lexikografisch dokumentiert, aber ohne Überprüfung der Richtigkeit der Angaben, der kontextuellen Relevanz oder der Zuverlässigkeit. Ein kritischeres Abwägen zwischen viel Information und zuverlässigen Angaben bleibt aus, und Nachschlagende müssen i. d. R. gewünschte Informationen selbst herausfiltern können.

> The wonderful thing about technology is that it can supply us with the volume of data that we need (and, increasingly, with the software for summarizing its salient features) in order to uncover and describe linguistic behaviour of this type. But the idea that the interpretive and 'synthetic' parts of lexicography can be automated to any significant degree seems to me unlikely and possibly misguided. (Rundell 2002: 152)

Die Stichproben zeigen, dass der Umgang mit dem Korpus zusammen mit den zugrunde gelegten Ansprüchen und Methodologien den entscheidenden Unterschied zwischen Quantität und Qualität eines Wörterbucheintrags darstellt. Wenn Atkins und Rundell (2008: 53) meinen, dass „objective evidence of language is a fundamental prerequisite for a reliable dictionary", dann ist es nötig hervorzuheben, dass zusätzlich zur Arbeitsgrundlage die Methoden entscheidend sind, um ein zuverlässiges Nachschlagewerk zu gestalten. Der Anspruch ein zuverlässiges Wörterbuch zu erarbeiten, kann nur dann eingelöst werden, wenn die verschiedenen zur Verfügung stehenden Verfahren und Methoden der Datengewinnung fortlaufend überprüft werden und die nötige lexikografische Auswertung und Interpretation von Korpusdaten nicht zunehmend in Widerspruch zur Erarbeitung eines korpusgestützten Wörterbuch gestellt wird.

Neben einer guten, in der Regel umfangreichen Datenbasis und geeigneten Methoden und Verfahren die Datenbasis explorativ zu ergründen, ist als drittes die linguistische Kompetenz eine wichtige Voraussetzung, gewonnene Daten zu interpretieren.

> For the forseeable future, tasks like this will be most effectively performed by a collaborative partnership of humans and machines. For we require not only high-quality data and cutting-edge software, but also that rare combination of editorial judgment, market knowledge, linguistic awareness, and good old-fashioned intuition […]. (Rundell 2002: 152–153)

Diesen Weg versucht seit einigen Jahren das Projekt *elexiko* (www.owid.de) bei der Erarbeitung eines korpusgestützten Internetwörterbuchs zu gehen. Das Projekt ist am Institut für Deutsche Sprache in Mannheim angesiedelt, und versucht neuen linguistischen und metalexikografischen Anforderungen gerecht zu werden, investigative Korpusansätze mit sich ergänzenden Korpusverfahren zu nutzen und die Auswertung der Daten kritisch mit lexikografischer Kompetenz zu komplementieren.

Es ist derzeit das einzige Projekt zur Beschreibung der deutschen Gegenwartssprache, das diesen Weg wählte. Diese Praxis erfordert ein hohes Maß an lexikografischer Erfahrung, um die Korpusresultate zu untersuchen, kontextuell zu interpretieren, Bedeutungen zu disambiguieren und Strukturen entsprechend zuzuordnen sowie die gewonnenen Erkenntnisse angemessen zu präsentieren und kohärent zu vernetzen (siehe Abbildung 3). Die lexikografischen Beschreibungen in *elexiko* sind aufgrund eines umfangreichen Korpus und aufgrund empirischer Untersuchungen objektiver und lückenloser, weil zueinander komplementäre Korpusverfahren zum Einsatz kommen und sie sind zuverlässiger, weil ein interpretatorisches Element der Klassifikation- und Filterleistung hinzukommt. Neben einem Korpusrecherche und -analysewerkzeug[4] sorgt der Einsatz eines maßgeschneiderten Managementtools[5] zusätzlich für konsistente Vernetzungen zwischen den Artikeln (z. B. bidirektional zwischen Synonymen). Die Inhalte der Artikelstruktur sind lexikografisch aufbereitet, auf ihre Konsistenz hin geprüft, ihre Vernetzung zu anderen Artikelstrukturen konsequent hergestellt. Nachteil dieser Wörterbucharbeit ist die äußerst zeitintensive Erarbeitung eines Wörterbuchartikels. Das liegt nicht ausschließlich an den genutzten Korpusverfahren und der linguistischen Bewertung der computerunterstützten Resultate, sondern auch an der umfangreichen inhaltlichen Informationsstruktur eines Eintrags.

Abbildung 3. Eingangsseite des Eintrags Mobilität in *elexiko* und Verweis auf weiterführende Kontextansicht.

Den derzeitigen lexikografischen Arbeitsaufwand vermögen gute, leistungsstarke Computerwerkzeuge heute bereits zu reduzieren. Tools, die linguistische Daten inhaltlich nach diversen Kriterien stärker ‚stromlinienförmig' gebündelt anbieten und den LexikografInnen damit eine zusätzliche Sortierung und Vorauswahl ermöglichen, werden vor allem in der englischen Lexikografie bereits genutzt.[6] Der Einsatz ähnlicher computationeller Verfahren könnte die Arbeit in *elexiko* optimieren.

5. Schlussbemerkungen

Elektronische Ressourcen spielen eine zunehmende Rolle und formen die Wörterbuchlandschaft. Gibt es in der Zukunft noch LexikografInnen? Diese Frage wird immer wieder aufgeworfen (z. B. Grefenstette 1998, Rundell 2002, Rundell/Kilgarriff 2011). Für die Erarbeitung einer lexikalischen Ressource mögen sie verzichtbar sein, für die Erarbeitung eines zuverlässigen Nachschlagewerks sind sie, ebenso wie ein geeignetes Korpus und die richtigen Methoden unerlässlich. Bereits Sinclair (1984: 4) weist darauf hin, dass die Lexikografie an der Schnittstelle zwischen Linguistik und Informationstechnologie agiert, dass aber ein dritter Faktor, die lexikografische Erfahrung mit ihren Prinzipien und Praktiken, eine ebenso relevante Größe ist. Das wird manchmal übersehen, wenn man die

Rolle des Dateninterpretatoren stillschweigend gerade auf jene überträgt, die i. d. R. eigentlich eine verlässliche Interpretation suchen, die NutzerInnnen. Insgesamt spielen je nach Ziel eines Wörterbuchs die einzelnen Faktoren möglicherweise unterschiedlich starke Rollen. In der deutschsprachigen monolingualen Lexikografie ist die Balance zwischen und das Miteinander der individuellen Ressourcen nicht immer ausgewogen.

Die englischsprachige Lexikografie beschreitet seit einiger Zeit den Weg der so genannten „Streamline-Tickboxlexikografie". Computergesteuerte Prozesse, die sprachliche Daten hinsichtlich unterschiedlicher linguistischer Kriterien automatisiert extrahieren, inhaltlich bündeln und stark vorsystematisiert anbieten, greifen mit kompetenzgesteuerter Selektion der gewonnenen Daten seitens der LexikografInnen ineinander und überführen die ausgewählten Daten ohne Zwischenschritte automatisch in das entsprechende Wörterbuchdokument an die ausgewiesenen Stellen (vgl. z. B. Projekt Dante - A lexical database for English). Die Visionen von künftigen Wörterbüchern (vgl. Rundell/Kilgarriff 2011) müssen und sollten neben sehr guten Korpora, guten Computertechnologien und komplexeren Automatisierungsprozessen die lexikografische Dateninterpretation nicht ausschließen, auch wenn neue Funktionen oder Rollen künftige LexikografInnen bei der Erarbeitung eines Wörterbuches erwarten.

> In this model, we envisage a change from the current situation, where the corpus software (some version of the word sketches) presents data to the lexicographer in (as we have seen) intelligently pre-digested form, to a new paradigm where the software selects what it believes to be relevant data and actually populates the appropriate fields in the dictionary database. In this way of working, the lexicographer's task changes from selecting and copying data from the software, to validating – in the dictionary writing system – the choices made by the computer. Having deleted or adjusted anything unwanted, the lexicographer then tidies up and completes the entry. (Rundell/Kilgarriff 2011: 278)

Entscheidend für die Wörterbuchqualität wird sein, wie gut LexikografInnen ihre veränderte Rolle bei der Entstehung von Wörterbüchern wahrnehmen werden. Ob es sie in ferner Zukunft noch geben wird, hängt aber auch davon ab, wie zuverlässig und authentisch NutzerInnen ihre Wörterbücher haben möchten.

Noten

[1] Etymologisches Wörterbuch des Deutschen von Wolfgang Pfeiffer (1997).
[2] Der Ausdruck *Ableben* wird zwar im Korpus aufgedeckt, aber aufgrund der Suchalgorithmen vom Computer als Hyponym interpretiert und dort falsch aufgelistet.
[3] Der Ausdruck *Tod* ist Bestandteil der Bedeutungserklärung von *Leben* und wird dort explizit als „Gegensatz" paraphrasiert. Als Antonym ist er jedoch nicht verzeichnet.
[4] In elexiko werden COSMAS II und die darin integrierte Software „Statistische Kollokationsanalyse und -clustering" genutzt (siehe Belica 1995).
[5] Vernetziko: Ein Vernetzungsmanagementtool, das im Rahmen von BZVelexiko entwickelt wurde und konsequentes bidirektionales Verlinken zwischen Relationselementen prüft. Siehe auch Meyer/Müller-Spitzer (2010) und Meyer (2011).
[6] Ein solches Verfahren stellt z. B. das Tool Sketch Engine (http://www.sketchengine.co.uk/) zur Verfügung, ein Korpusrecherchesystem, das inhaltlich die Verwendungsmöglichkeiten eines Ausdrucks nach grammatischen und kollokationalen Kriterien gebündelt zusammenfasst.

Literatur

A. Korpora, Tools und Nachschlagewerke (Stand 12. März 2012)

***BZVelexiko*.** Benutzeradaptive Zugänge und Vernetzungen in elexiko. http://www1.ids-mannheim.de/lexik/BZVelexiko/.

***COSMAS II*.** Corpus Search, Management and Analysis System. http://www.ids-mannheim.de/cosmas2/uebersicht.html.

***Dante*.** A lexical database for English. http://www.webdante.com/index.html.

***DeReKo*.** Deutsches Referenz Korpus. http://www.ids-mannheim.de/kl/.

***Duden-GWDS*.** Das große Wörterbuch der deutschen Sprache in 10 Bänden auf CD-ROM. (Third edition) Bibliographisches Institut: Mannheim, 2000.

***Duden online*.** www.duden.de.

***DWDS*.** Digitales Wörterbuch der Deutschen Sprache. http://www.dwds.de/.

***elexiko*.** www.elexiko.de/ oder www.owid.de.

***Open Thesaurus*.** http://www.openthesaurus.de.

Pfeifer, W. 1997. *Etymologisches Wörterbuch des Deutschen*. (Third edition) München: Dt. Taschenbuch-Verlag.

***Sketch Engine*.** http://www.sketchengine.co.uk/?page=Website/SketchEngine.

***Wortschatzlexikon, Wortschatzportal*.** http://wortschatz.uni-leipzig.de/.

B. Weitere Literatur

Atkins, S. and M. Rundell 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Belica, C. 1995. *Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethode*. Mannheim: Institut für Deutsche Sprache. (http://corpora.ids-mannheim.de/.)

Grefenstette, G. 1998. 'The Future of Linguistics and Lexicographers: Will there be Lexicographers in the year 3000?' In T. Fontenelle et al. (Hrsg.), *Proceedings of the 8th EURALEX International Congress on Lexicography* in Liege, Belgium. Liege: English and Dutch Departments, University of Liege, 25–41.

Hanks, P. 1990. 'Evidence and Intuition in Lexicography.' In J. Tomaszczyk and B. Lewandowska-Tomaszczyk (Hrsg.), *Meaning and Lexicography*. Amsterdam/Philadelphia: Benjamins, 31–41.

Heid, U. 2010. 'Computergestützte Lexikographie und Terminologie.' In K. U. Carstensen, Ch. Ebert, C. Ebert, S. Jekat, H. Langer and R. Klabunde (Hrsg.), *Computerlinguistik und Sprachtechnologie: Eine Einführung*. (3. Auflage) Heidelberg: Spektrum Akademischer Verlag, 566–575.

Hunston, S. 2002. *Corpora and Applied Linguistics*. Cambridge: Cambridge University Press.

Meyer, P. 2011. 'vernetziko: A Management Tool for the Lexicographer's Workbench.' In I. Kosem and K. Kosem (eds.) *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011. 10-12 November 2011. Bled, Slovenia*. Ljubljana: Trojina. (http://www.trojina.si/elex2011/Vsebine/proceedings/eLex2011-25.pdf).

Meyer, P. and C. Müller-Spitzer 2010. 'Consistency of Sense Relations in a Lexicographic Context.' In V. Barbu Mititelu, V. Pekar and E. Barbu (Hrsg.), *Proceedings of the Workshop Semantic Relations. Theory and Applications*, 18 May 2010, at the International Conference on Language Resources and Evaluation (LREC) 2010, Malta, 37–47. (http://www.lrec-conf.org/proceedings/lrec2010/workshops/W9.pdf).

Müller-Spitzer, C., Koplenig, A. and A. Töpel 2011. 'What Makes a Good Online Dictionary? – Empirical Insights from an Interdisciplinary Research Project.' In I. Kosem and K. Kosem (eds.) *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011. 10-12 November 2011. Bled, Slovenia*. Ljubljana: Trojina

Rundell, M. 2002. 'Good Old-fashioned Lexicography: Human Judgment and the Limits of Automation.' In M.-H. Corréard (Hrsg.), *Lexicography and Natural Language Processing, A Festschrift in Honour of B. T. S. Atkins*, Grenoble: EURALEX, 138–155.

Rundell, M. and A. Kilgarriff 2011. 'Automating the creation of dictionaries: where will it all end?', In F. Meunier, S. De Cock, G. Gilquin and M. Paquot (Hrsg.), *A Taste for Corpora. In honour of Sylviane Granger*. Amsterdam/Philadelphia: John Benjamins, 257–281.

Sinclair, J. 1984. 'Lexicography as an academic subject.' In R. R. K. Hartmann (Hrsg.), *LEXeter '83 Proceedings*. Tübingen*:* Max Niemeyer, 3–12. (Lexicographica Series Maior 2).

Sinclair, J. 1997. 'Corpus Evidence in Language Description.' In A. Wichmann, St. Fligelstone, T. McEnery and G. Knowles (Hrsg.), *Teaching and Language Corpora*. London/New York: Longman, 27–39.

Storjohann, P. 2005. 'Corpus-driven vs. corpus-based approach to the study of relational patterns.' In *Proceedings of the Corpus Linguistics Conference 2005* in Birmingham. (http://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx).

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.