Vom Textkorpus zur Datenbank: Probleme der Systematisierung der Daten bei der Erarbeitung eines Wörterbuchs

Irina Kruashvili
Staatliche Sokhumi-Universität, Tbilisi, Georgien
irina555k@yandex.ru

Die Anforderung an die moderne Lexikographie ist es, ein Nachschlagewerk sowohl in Buch- als auch elektronischer Form zu schreiben. Der Beitrag geht der Frage nach, vor welche Probleme man sich hinsichtlich der Systematisierung der Daten und ihrer Eintragung in die Datenbank gestellt sieht. Er beschreibt den Weg, auf dem die Wörter vom Textkorpus in die Datenbank gelangen.

Schlüsselwörter: **Datenbank, Informationssystem, Lemmazeichen, Schreibvariante, Lexem**.

1. Einleitung

Die heutige georgische Lexikographie ist, historisch gesehen, in eine merkwürdige Lage geraten: Einerseits erscheinen dilettantisch geschriebene Wörterbücher mittleren Umfangs in effektvollen Umschlägen, deren Analyse eine ungenügende theoretische Ausbildung der Verfasser an vielen Beispielen sowohl in der Makro- als auch in der Mikrostruktur enthüllt. Auf dem Markt erscheinen weiterhin jahreleng dieselben Wörterbücher, die – kurz charakterisiert – immer wieder aufs neue ausgeflickt und graphisch ausgebessert werden. Auch ihre ursprüngliche Konzeption kann der neueren lexikographischen Praxis nicht mehr standhalten. Andererseits erfahren wir aus dem Internet und aus der deutsch- und englischsprachigen Fachliteratur dank umfassenden Arbeitsberichten von computer- bzw. corporagestützten lexikographischen Projekten, deren Informationsangebot die gängigen Wörterbücher übertrifft. Um aus dieser Situation einen Ausweg zu finden, ist es nötig, sich sowohl von europäischen ein- als auch zweisprachigen Wörterbüchern inspirieren zu lassen und die schnelle Entwicklung der technischen Speicherung und Darstellung ernst zu nehmen.

2. Die Datenbank

Unter Berücksichtigung des oft beklagten Zustandes der georgischen Lexikographie entsteht die Notwendigkeit, ein neues großes deutsches Wörterbuch zu erarbeiten. Im Laufe der Zeit änderte sich die Vorstellung von der Ergebnisform: Nicht nur ein Wörterbuch soll publiziert werden, sondern unsere Ergebnisse der Erfassung, Beschreibung und Dokumentation der Wörter sollen zunächst in das Informationssystem eingebracht werden. Die Anforderung an die moderne Lexikographie ist es, ein Nachschlagewerk sowohl in Buch- als auch elektronischer Form zu schreiben. Zu seinem Adressatenkreis gehören natürlich vor allem Akademiker, Hochschulstudenten und Übersetzer. Das Wörterbuch soll nicht nur den lebendigen deutschen Sprachgebrauch, nicht nur den Wortschatzkern der wichtigsten wissenschaftlichen Bereiche abdecken, sondern auch die Wörter auflisten, die für diejenigen unentbehrlich sind, die sich mit der deutschen Literatur und Geschichte befassen.

Der erste Schritt besteht darin, ein lexikalisch-lexikologisches korpusbasiertes Informationssystem zu entwickeln. Eine der DV-Komponenten, aus denen Informationssystem besteht, ist eine Datenbank. In dieser objektrelationalen Datenbank sollen die Ergebnisse aller wortschatzbezogenen Projekte sowie mittelfristig auch Ergebnisse externer wissenschaftlicher Wortschatzforschungen abgelegt, gebündelt und in Beziehung zueinander gebracht werden, um so neues linguistisches Wissen möglich zu machen.

Die Datentypen ordnen sich folgenden drei Informationsdimensionen eines Suchwortes zu:

1. Schreibung und Aussprache
2. Bedeutung und Verwendung

3. Grammatik

Eine Ausnahme bildet diesbezüglich die Lemmazeichengestaltangabe. Die Lemmazeichengestaltangabe kann die Verfasser des Wörterbuchs in orthographischer Hinsicht vor Probleme stellen. In den Texten des Korpus trifft man nicht zufällig auf unterschiedliche Schreibungen ein und desselben Lexems, hat sich doch bei manchen Wörtern die Schreibnorm noch nicht gefestigt. Während Komposita, deren Unmittelbare Konstituenten schon seit langem etablierte Lexeme sind, häufig in zwei Schreibvarianten – einer ohne Bindestrich (z. B. „Schlüssellochchirurgie“) und einer mit Bindestrich (z. B. „Schlüsselloch-Chirurgie“) – belegt sind, können sich bei manchen Fremdwörtern auffallend viele Schreibvarianten finden, die einen unterschiedlichen Grad der Integration in die deutsche Schreibnorm zeigen: „Couchpotato“ („jemand, der gern fernsieht und dabei auf der Couch sitzt und Salzgebäck, Süßigkeiten isst“) mit den Varianten „Couch-Potato“, „Couch-potato“, „couch-potato“„Couch Potato“, „Couch potato“, „couch potato“.

Natürlich erfolgt die Lemmatisierung der Wörter entsprechend den Regeln der neuen deutschen Rechtschreibung, von denen die zur Schreibung mit Bindestrich und zur Getrennt- und Zusammenschreibung für uns von besonderem Interesse sind. Probleme können sich für den Lemmaansatz besonders von Fremdwörtern dann ergeben, wenn Variantenschreibung zugelassen ist. Werden Haupt- und Nebenvariante unterschieden, wird die Hauptvariante als Lemmazeichengestaltangabe erscheinen. Sind aber Fremdwörter bei Zusammenschreibung sehr unübersichtlich, sollte es erwogen werden, bei Vorhandensein normgerechter gleichberechtigter Schreibvarianten für den Lemmaansatz die Variante zu wählen, die dem Leser eine Hilfe für das Verständnis an die Hand gibt, z. B. Bindestrich-Schreibung: „Shareholder-Value“, Getrenntschreibung: „Golden Goal“, „Electronic Banking“.

4. Angabe zu den Schreibvarianten, zur Silbentrennung und zur Aussprache

Es wäre zweckmäßig, alle in den Texten belegten Schreibvarianten eines Wortes zu verzeichnen. Dabei sollen die den Regeln entsprechenden von den nicht den Regeln entsprechenden Varianten abgehoben werden. Anschließend folgt die Angabe der Silbentrennung.

Für jedes Wort soll die Aussprache mit den Zeichen der International Phonetic Association (IPA) wiedergegeben werden. Besonders bei Fremdwörtern sieht man sich dabei vor einem großen Problem, da meist nur Belege aus verschriftlichten Texten zur Verfügung stehen, so dass man nicht weiß, ob und – wenn ja – in welchem Grad diese Fremdwörter im Deutschen phonetisch eingebürgert wurden. Soll man sich z. B. für „Golden Goal“ an der in Duden – Oxford Großwörterbuch Englisch [1] für „golden“ und „goal“ angegebenen Aussprache [gəʊldn] bzw. [gəʊl], an der in Duden – Universalwörterbuch [2] für „Golden Goal“ angegebenen Aussprache [gəʊldən gəʊl] oder an der Aussprache [gouldəngoul] orientieren, die der in Wahrig: Fremdwörterlexikon [3] für „Goldengoal“ angegebenen entspräche, oder aber mehrere Varianten berücksichtigen [4], [5], und wären mit ihnen wirklich alle Aussprachevarianten erfasst? [6], [7].

5. Semantische Angaben

Jedes Wort erhält eine semantische Paraphrasenangabe, für die die Belegtexte, in denen ein Wort nicht selten erklärt wird, eine Hilfe an die Hand geben können, so, wenn es z. B. in einem Beleg für „Flyer“ heißt:

„Seine Vorliebe gilt den „Underground-Unternehmungen“, Partys, die in ungewöhnlicher Umgebung stattfinden und deren Termine nur über „Flyer“ zu erfahren sind, jene gestylten Handzettel, die nachts in Kneipen und Diskotheken verteilt werden“ (Frankfurter Allgemeine Zeitung, 05.05.1995).

Ein solcher Beleg kann als Definitionsbeleg der semantischen Paraphrasenangabe – „Flyer“: „häufig computergrafisch gestalteter Handzettel, mit dem für (Szene)partys o. ä. geworben wird“ – hinzugefügt werden. Für jedes Wort sollten Belegbeispiele angeführt werden, die in semantischer Hinsicht möglichst sprechend sein sollten. Besonders sollten auch solche Belegbeispiele zitiert werden, in denen durch Markierungen auf den Neuheitscharakter des betreffenden Lexems, der betreffenden Bedeutung hingewiesen wird, z. B. durch Anführungszeichen, durch Kursivdruck, durch Hinzufügen von „so genannt“ oder durch eine beigegebene Bedeutungserläuterung. Auf die Funktion der Markierungen sollte man in einem Kommentar eingehen. Zusätzlich zu den Belegbeispielen könnte man für ein Wort Syntagmen verzeichnen: „Flyer“ – „Flyer verteilen“.

Es wäre wünschenswert, dass für jedes Fremdwort angegeben wird, seit wann es belegt ist. Nur ausnahmsweise kann man ziemlich genau den Zeitpunkt des Aufkommens eines Wortes bestimmen. So wissen wir, dass „Euro“ als Bezeichnung für die neue europäische Währung auf dem Gipfeltreffen der Regierungschefs der 15 Mitgliedsländer der Europäischen Union im Dezember 1995 festgelegt wurde.

Natürlich gibt es keine Möglichkeit, für die Fremdwörter-Neologismen Erstbelege zu geben, man sollte aber bemüht sein, jeweils einen möglichst frühen Beleg anzuführen und durch die Belegauswahl auch das zeitliche Kontinuum der Verbreitung zu dokumentieren.

Zu den Wörtern können pragmatische Angaben gemacht werden. Diese können z. B. die Zuordnung der Wörter zu einem Sach-/Fachbereich, zu einer Fachsprache betreffen oder aus Hinweisen auf die Gebundenheit ihrer Verwendung an spezifische Textsorten oder an bestimmte Situationen bestehen.

6. Grammatische Angaben

Für Substantive, die Singular und Plural bilden, sollen das Genus, der Genitiv Singular und der Nominativ Plural angegeben werden [8]. Handelt es sich bei einem Substantiv um ein Singulariatantum bzw. um ein Pluraliatantum, soll es als solches charakterisiert werden. Das Singulariatantum erhält zudem die Genusangabe und die Angabe des Genitivs Singular, das Pluraliatantum die des Nominativs Plural. Bei jeder Angabe des Nominativs Plural soll explizit vermerkt werden, ob Umlaut vorhanden ist oder nicht.

Dass besonders bei substantivischen Fremdwörtern grammatische Varianten auftreten können, sei an folgenden Beispielen verdeutlicht:

Varianten beim Genus: „Event“ – Neutrum, Maskulinum [9].

„Ein richtiges Event war es allerdings dennoch nicht“ (Die Tageszeitung, 04.09.1998). Aber: „Dennoch sollte dieser Event eine Premiere sein“ (Die Tageszeitung, 24.12.1996).

In einem Kommentar soll darauf hingewiesen werden, dass die Mehrzahl der Belege fur „Event“, in denen das Genus erkennbar ist, das Lexem als Neutrum (vermutlich nach seiner lexikalischen Entsprechung „das Ereignis“) zeigt.

Varianten beim Genitiv Singular: „Car-Sharing“ – Gen. Sg. „Car-Sharing“, „Car-Sharings“.

„Berlin schmückt sich mit der Idee des Car-Sharing“ (Die Tageszeitung, 04.03.1997). Aber: „Die Idee des Car-Sharings [wird] erläutert“ (Mannheimer Morgen, 17.04.1999).

Bei den zahlreichen Fremdwörtern mit neutralem Genus, bei denen es sich aus Sicht der englischen Wortbildung um Verbalsubstantive mit dem Suffix „-ing“ oder um Komposita mit einem solchen Verbalsubstantiv als Grundwort handelt, treten Schwankungen auf zwischen dem endungslosen Genitiv Singular, der der englischen Flexion entspricht, und dem Genitiv Singular mit der Endung „-s“, der der deutschen Flexion entspricht. Deswegen soll bei den entsprechenden Fremdwörtern in der Regel sowohl der Genitiv mit der Endung „-s“ als auch der endungslose Genitiv angegeben werden.

Durch die Angaben zur Wortbildung sollen die Wörter im Hinblick auf ihre Wortbildungsart charakterisiert werden. Bei den Wörtern, für die eine solche Charakterisierung möglich ist, handelt es sich in der Regel um die Charakterisierung als Kompositum, Ableitung und Kurzwort. Als Beispiele seien genannt:

Kompositum – Bei einem Wort, der als Kompositum charakterisiert wird, sollen, sofern es sich bei dessen Unmittelbaren Konstituenten um freie Grundmorpheme bzw. freie Morphemkonstruktionen handelt, diese angeführt und ihre Wortart verzeichnet werden. Außerdem soll angegeben werden, ob ein Fugenelement vorhanden ist und, wenn ja, welches, z. B. „Abonnementsfernsehen“: Kompositum aus „Abonnement“ (Substantiv) + Fuge „-s-„ + „Fernsehen“ (Substantiv). Zusätzlich können Aussagen zu der semantischen Binnenrelation zwischen den Unmittelbaren Konstituenten gemacht werden.

Ableitung – Bei einem Wort, das als Ableitung charakterisiert wird, sollen, sofern es sich bei der Derivationsbasis um ein freies Grundmorphem bzw. eine freie Morphemkonstruktion und beim Derivationsaffix um ein Suffix handelt, diese angeführt und die Wortart des freien Grundmorphems bzw. der freien Morphemkonstruktion verzeichnet werden. Außerdem soll angegeben werden, ob ein Fugenelement vorhanden ist und, wenn ja, welches, und ob Umlaut eingetreten ist, z. B. „ostig“: Ableitung aus „Osten“ (Substantiv) mit Tilgung von „-en“ + Suffix „-ig“, Umlaut: nein.

Fremdwörter, die als fertiges Wortbildungsprodukt aus dem Englischen ins Deutsche gelangt sind (z. B. „Shareholder-Value“, „Flyer“), erhalten keine Charakterisierung z. B. als Kompositum oder als Ableitung, da die Angaben zur Wortbildung bei den im Deutschen gebräuchlichen Fremdwörtern auf der deutschen, nicht aber auf der englischen Wortbildung basieren. Es sollen aber bei solchen Fremdwörtern die ihnen zugrunde liegenden englischen Konstituenten mit ihren deutschen Entsprechungen in die Datenbank eingetragen werden, z. B. „Shareholder-Value“: engl. „shareholder“ – „Aktionär“, engl. “value” – „Wert“; “Flyer”: engl. “to fly” – „fliegen“.

Die Wortbildungsproduktivität eines Lexems soll durch die Aufzählung von zu ihm gebildeten Komposita und/oder Ableitungen angedeutet werden, z. B. „Flyer“: „Party-Flyer“, „Techno-Flyer“.

7. Zusammenfassung

Wir haben versucht, den Weg zu beschreiben, auf dem die Wörter vom Textkorpus in die Datenbank gelangen. Unsere Ausführungen wollen keinen Anspruch auf Endgültigkeit erheben. Es wird wohl noch einige Zeit vergehen, ehe man alle Probleme gelöst hat, vor die man sich hinsichtlich der Systematisierung der Daten und ihrer Eintragung in die Datenbank gestellt sieht.

Literatur

[1] Duden, „Oxford Großwörterbuch Englisch, Englisch-Deutsch, Deutsch-Englisch“, Hg. von der Dudenredaktion und Oxford University Press. Redaktionelle Leitung: Werner Scholze-Stubenrecht, John B. Sykes. Mannheim, Bibliographisches Institut, 2005.

[2] Duden, „Deutsches Universalwörterbuch“, 7., überarbeitete und erweiterte Auflage. Mannheim, Bibliographisches Institut, 2011.

[3] R. Wahrig, „Fremdwörterlexikon“, Hg. von Renate Wahrig-Burfeind, Oliver Mingers, 6., vollständig neu bearbeitete und aktualisierte Auflage. Gütersloh/München, Wissen Media Verlag (vormals Bertelsmann Lexikon Verlag), 2007.

[4] Duden, „Das Große Fremdwörterbuch. Herkunft und Bedeutung der Fremdwörter“, Hg. von Dudenredaktion, 4., aktualisierte Auflage.

Mannheim, Bibliographisches Institut & E. A. Brockhaus, 2007.
[5] Duden, “Die deutsche Rechtschreibung“, Der Duden in 12 Bänden. Das umfassende Standardwerk auf der Grundlage der neuen amtlichen Rechtschreibregeln, Bd. 1. Hg. von der Dudenredaktion, 24. völlig neu bearbeitete und erweiterte Auflage. Mannheim, Bibliographisches Institut & E. A. Brockhaus, 2006.
[6] R. Wahrig, “Deutsches Wörterbuch”, Hg. von Renate Wahrig-Burfeind, 8., vollständig neu bearbeitete und aktualisierte Auflage. Gütersloh/München, Wissen Media Verlag (vormals Bertelsmann Lexikon Verlag), 2006.
[7] Duden, „Das Große Wörterbuch der deutschen Sprache in zehn Bänden“, Hg. vom Wissenschaftlichen Rat der Dudenredaktion. Berlin, Bibliographisches Institut, 2002.
[8] Langenscheidt, “Langenscheidts Großwör-terbuch. Deutsch als Fremdsprache”, Hg. von Dieter Götz, Günther Haensch, Hans Wellmann. Berlin, München, Wien, Zürich, New York, 2007.
[9] Bertelsmann, „Die deutsche Rechtschrei-bung“, verfasst von Ursula Hermann, völlig neu bearb. und erweitert von Lutz Götze. Gütersloh/München, 1999.