

Zur Anwendung von korpusbasierten Sprachdaten in der pädagogischen Lexikographie

Dr. Milka Enčeva
Universität Maribor/Slowenien

Ziel des Beitrags

- die bestehenden elektronischen Wörterbücher weit hinter den Möglichkeiten des elektronischen Mediums (Müller-Spitzer 2003: 141)
- Wie ist die Situation 12 Jahre später?
- im Hinblick auf zweisprachige Online Lernerwörterbücher (OLWB)

Inhaltsübersicht

1. Lerner- (Benutzer-)gruppen
2. Lemmaselektion
3. Kollokationen
4. Fazit

1. Lerner-(Benutzer-)Gruppen

- in den 90-er Jahren des 20. Jhs. – Erstellung von einsprachigen DaF-LWB für die gesamte Gruppe der Deutschlernenden
- Differenzierung der Lernergruppen nach dem Niveau ihrer Fremdsprachenkenntnisse nötig

Zum Vergleich

1. In Pons Online Wörterbuch keine Differenzierung der Lernergruppen nach dem Niveau der Fremdsprachenkenntnisse
2. In ELDIT (Elektronische Lernerwörterbuch Deutsch-Italienisch) eine Lernergruppe - Anfänger bis leicht Fortgeschrittene (A1, A2 und B1 Niveau)
(vgl. Abel 2008: 177)
 - problematisch:
 - die Behandlung von A1/A2 und B1 als eine Gruppe
(vgl. Europarat/GER 2001)

Vorschlag

- zwei selbstständige Lernergruppen bzw. zwei OLWB als eine Datenbank:
 - OLWB1 für das Sprachniveau A1 – A2 (elementare Sprecher) und
 - OLWB 2 für das B1 – B2 (selbstständige Sprecher),
- für C1 – C2 (kompetente Sprecher) - kein OLWB, weil: die Lerner ein einsprachiges LWB oder ein zweisprachiges ch der Allgemeinsprache benutzen können (vgl. Glaboniat u. a. 2005: 68).

2. Lemmaselektion

- Vielzahl von Grund- und Aufbauwortschätzen zur deutschen Sprache

problematisch:

- nicht auf transparenter, empirischer Basis zusammengestellt
- beruhen auf individuellen Kenntnissen und Erfahrungen von Lehrenden oder Wissenschaftlern

- Vermeidung einer mechanischen Übernahme von Wortschatzlisten (z.B. der Liste von Profile Deutsch 2005),
- Berücksichtigung neuer Häufigkeitslisten des Grund- und Aufbauwortschatzes der deutschen Sprache (z.B. Häufigkeitsliste des Herder-Instituts - Jones/Tschirner 2006).

- Kempckes (1992) Schlussfolgerung
„[...]daß es *den* Grundwortschatz nicht gibt.“
- Behandlung einer Teilmenge des Gesamtwortschatzes durch eigene Selektion und je nach Zweck und Adressatengruppe(n) als Grundwortschatz
(Tschirner 2005, 2008, Lübke 2008, Bubenhofer 2012, Lange u.a. 2015)

Ansätze zur Bestimmung des Grundwortschatzes

1. Kommunikativ-pragmatischer Ansatz:

- von typischen kommunikativen Situationen (Alltagssituationen) und Sprechintentionen ausgehend
z.B. Kontaktschwelle Deutsch (1980),
Zertifikat Deutsch (1972, 1992) des Goethe-Instituts
die deutsche Fassung des GER in Profile Deutsch (2005)

Problematisch:

- die Begriffe *Alltag* und *Situationen* weder theoretisch noch empirisch fundiert
- unklar die Selektionsverfahren zur Ermittlung der Lexeme

2. Lexikographischer Ansatz:

- auf der Grundlage von Sekundärquellen (Wortschatzsammlungen und bestehenden Wörterbüchern)

z.B. Schnörch (2002), Haderlein (2008)

Problematisch:

- die Quellen,
- Intransparenz der Selektionskriterien.

3. Frequenzorientierter Ansatz:

- über Häufigkeitszählungen zu einer Wortschatzbegrenzung

z.B. Jones/Tschirner (2006)

Tschirner (2005, 2008)

problematisch:

- kein 1:1 Verhältnis zwischen hoher Frequenz und kommunikativer Wichtigkeit
- keine Berücksichtigung der potentiellen Benutzer (Lerner), der Kommunikationssituationen und der relevanten Handlungsmuster.

	Frequenz-orientierter Ansatz	Kommunikativ-pragmatischer Ansatz	Lexikographischer Ansatz
Vertreter	Jones/Tschirner (2006), Tschirner (2005, 2008)	Kontaktsprache Deutsch (1980) Goethe Institut (1972, 1992) Glaboniat u.a. (2005)	Schnörch (2002) Haderlein (2008)
Datenbasis	primäre Quellen (Korpora, Texte)	selektive oder sporadische Nutzung empirischer Daten	sekundäre Quellen (Wörterbücher, Wortschatz- sammlungen)
Haupt- kriterium der Selektion	Häufigkeit	kommunikative Relevanz	Schnittmengenbildung oder Bündel aus statistischen Maßen + Introspektion
Nachteile	- hohe Frequenz \neq kommunikative Wichtigkeit; - unberücksichtigt: ▪ potentielle Benutzer, ▪ Kommunikationssitua- tionen, ▪ relevante Handlungsmuster	- keine theoretische Fundierung und empirische Absicherung der Kategorien (Situationen, Themen, Sprechakte); - Intransparenz der Selektionskriterien und Wortschatzzuordnung	- keine primären Quellen, - Selektionskriterien

Der mehrdimensionale Frequenzbegriff

- relative Frequenz
- temporale Stabilität – über einen längeren Zeitraum gleichmäßiges häufiges Auftreten eines Lemmas
- themenübergreifende Stabilität – das gleichmäßig häufige Auftreten eines Lemmas in unterschiedlichen Sachbereichen
- Produktivität – Wortbildungspotenz

(vgl. Okamura, Saburo / Willi Lange / JoachimScharloth 2012)

Wahl des Korpus

“...fast alle Korpora der deutschen Sprache, auf denen Häufigkeitswörterbücher und Grund- und Aufbauwortschätze basieren, [sind] veraltet ...und viele davon wenig repräsentativ für die deutsche Sprache in ihrer Gesamtheit im gesamten deutschsprachigen Raum“.

(Tschirner 2005: 5)

Exkurs:

- das erste Häufigkeitswörterbuch der deutschen Sprache von Kaeding (1897) erstellt
- 1975 von Ortman überarbeitet und als *Hochfrequente Wortformen I und II* herausgegeben
- Ortmanns *Hochfrequente Wortformen* – Grundlage für die Wortschatzliste des Goethe- Instituts für das Zertifikat Deutsch (1992)
- die Wortschatzliste des Goethe-Instituts in Profile Deutsch (2002) übernommen

Das Herder/BYU-Korpus

- Korpusumfang von 4,2 Mio. Wortformen
 - ein aktuelles, ausgewogenes Korpus - Zeitungstexte, Sach- und Fachtexte, literarische Texte, Gebrauchstexte und gesprochene Sprache
 - ein für die deutschsprachigen Länder repräsentatives Korpus - das Verhältnis zwischen deutschen, österreichischen und schweizer Texten ca. 70 : 20 : 10
 - das daraus resultierende Häufigkeitswörterbuch - 4000 Lexeme
- (Tschirner 2005: 6)

Das Forschungsprojekt „Korpusgeleiteter Grund- und Aufbauwortschatz Deutsch“

- Korpus - mehr als 100 Millionen Wörter
- statt Repräsentativität und Ausgewogenheit - prototypische Kommunikationskonstellationen:
 - mehrfachadressierte und konzeptionell schriftliche Texte - Zeitungen
 - persönlich adressierte und konzeptionell mündliche Texte - Online-Diskussionsforen

Vorschlag

- mehrdimensionaler Frequenzbegriff
- prototypische Kommunikationskonstellationen
- Repräsentativität –
- OLWB 1 – 2000 Lexeme
- OLWB 2 – 4000 Lexeme

Kollokationen in zweisprachigen OLWB

- OLWB =idealer Ort zur Beschreibung von im lexikographischen und sprachdidaktischen Sinn relevanten Kollokationen (Herbst-Klotz 2003)
- Kollokationsforschung in der Germanistik und in der Slowenistik und Bulgaristik:

Slowenistik:

- vor allem in kontrastiven Arbeiten Deutsch – Slowenisch und Englisch- Slowenisch (vgl. Gruntar- Ermol (2003), Gorjanc/Jurko 2004)

Bulgaristik:

- vor allem in kontrastiven Arbeiten Deutsch-Bulgarisch und Englisch-Bulgarisch (Baschewa 2010, Petkova-Kessanlis 2010)

- computergestützte kontrastive Sprachforschung – wichtiger Beitrag zur besseren Unterstützung kleinerer europäischer Sprachen durch Sprachtechnologie

- Die Studie „Europe’s Languages in the Digital Age“ von META-NET (2012):
 - Stand der Sprachtechnologie für 30 europäische Sprachen
 - Bewertung der sprachtechnologischen Unterstützung in vier Anwendungsbereichen:
 - automatische Übersetzung,
 - Erkennung und Generierung gesprochener Sprache,
 - Textanalyse und
 - Sprachressourcen.

Sprach- und Textressourcen

ausgezeichnete Unterstützung	gute Unterstützung	moderate Unterstützung	fragmentarische Unterstützung	schwache/ keine Unterstützung
	Englisch	Deutsch Französisch Italienisch Niederländisch Polnisch Schwedisch Spanisch Tschechisch Ungarisch	Baskisch, Bulgarisch Dänisch Estnisch Finnisch Griechisch Katalanisch Kroatisch Norwegisch Portugiesisch Rumänisch Serbisch Slowakisch Slowenisch	Irish Isländisch Litauisch Lettisch Maltesisch Welsch

<http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison> (Abruf am 3.10.2015)

Vorschlag von Möhring (2011) für eine systematische Wörterbuchanalyse:

- Berücksichtigung von semantischen, pragmatischen und korpuslinguistischen Kriterien
- Identifikation von statistisch signifikanten Kollokationspartnern – DWDS Kernkorpus
- manuelle Überprüfung und Selektion
- Liste mit 56 hochfrequenten lexikalischen Kollokationen (zu 26 Ausgangswörtern)
- das Vorkommen dieser Kollokationen in GWB, LDaF, PDaF, WAR analysiert

Vorkommen der Kollokationen in PONS Online Wörterbuch Deutsch-Slowenisch und Deutsch-Polnisch:

Kollokationen	Deutsch-Slowenisch	Deutsch-Polnisch
vorhanden	26/56	36/56
nicht vorhanden	30/56	20/56
in beiden Wörterbüchern vorhanden	25/56	
in beiden Wörterbüchern nicht vorhanden	15/56	

Kollokationsidentifikation:

- Datenerhebung mittels korpusanalytischer Methoden,
- manuelle Auswertung der Daten unter pragmatisch-semantischen Gesichtspunkten.

4. Fazit

- neue Korpora als Ausgangsbasis für die Festlegung eines Grundwortschatzes nötig - neben Frequenz prototypische Kommunikationskonstellationen zu berücksichtigen
- Lemmaselektion - mehrdimensionaler Frequenzbegriff
- kontrastive Kollokationsforschung als Beitrag zur Förderung der Kollokationsforschung kleinerer Sprachen
- korpuslinguistische Datenerhebung in Kombination mit Auswertung der Daten unter semantischem und pragmatischem Aspekt

- die Aussage von Müller-Spitzer (2003) immer noch aktuell

Aber:

- An der Erstellung neuer, besserer, moderner elektronischer Wörterbücher wurde und wird gearbeitet.

Wörterbücher

1. Pons-Online Wörterbuch (<http://de.pons.com/> Abruf am 3.10.2015)
2. ELDIT (Elektronisches Lernerwörterbuch Deutsch-Italienisch) (<http://eldit.eurac.edu/> Abruf am 3.10.2015)

Literatur

1. Abel, Andrea (2008): ELDIT (Elektronisches Lernerwörterbuch Deutsch – Italienisch und elexiko: ein Vergleich. In: Klosa, Anette (Hg.): *Lexikografische Portale im Internet*. (=OPAL Sonderheft 1/2008), 175-189.
2. Baldegger, Markus/ Martin Müller/Günther Schneider (1980): *Kontaktschwelle Deutsch als Fremdsprache*. Langenscheidt: Berlin u.a.
3. Deutscher Volkshochschulverband/Goethe-Institut (Hrsg.) (1992): *Das Zertifikat Deutsch als Fremdsprache*. Bonn u.a.

4. Baschewa, Emilia (2010): Zur lexikografischen Bearbeitung von Kollokationen in allgemeinen zweisprachigen Wörterbüchern Bulgarisch-Deutsch. In: Durčo, Peter (Hrsg.): *Feste Wortverbindungen und Lexikographie. Kolloquium zur Lexikographie und Wörterbuchforschung*. De Gruyter: Berlin, New York, 9-22.
5. Bubenhofer, Noah (2012): Lehrwerke und Referenzwortschätze. Der Nutzen frequenzbasierter Grundwortschätze. In: Okamura, Saburo/Willi Lange/Joachim Scharloth (Hrsg.): *Grundwortschatz Deutsch: lexikografische und fremdsprachendidaktische Perspektiven*. JGG :Tokyo. 13-27. (=Studienreihe der Japanischen Gesellschaft für Germanistik 088).
6. Europarat (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen (GERS)*. Modern Languages Division. Berlin: Langenscheidt.
7. Glaboniat, Manuela u. a. (2005):). *Profile deutsch. Langenscheidt. Lernzielbestimmung, Kann Beschreibungen und kommunikative Mittel für die Niveaustufen A1, A2, B1, B2, C1 und C2 des »Gemeinsamen europäischen Referenzrahmens für Sprachen«*. Berlin: Langenscheidt.

8. Gruntar- Ermol, Ada (2013): Die Bedeutung fachsprachlicher Kollokationen bei der Textproduktion. In : Fabčić, Melanija , Sabine Fiedler, Joanna Szerszunowicz (Hrsg.): *Phraseologie im interlingualen und interkulturellen Kontakt = Phraseology in interlingual and intercultural contact*, (Mednarodna knjižna zbirka Zora, 95). Mednarodna založba Oddelka za slovanske jezike in književnosti, Maribor, 93-109.
9. Gorjanc, Vojko/Primož Jurko (2004): Kolokacije in učenje tujega jezika. *Jezik in slovstvo* 49, 3/4, 49- 62.
10. Jones, Randall L./Tschirner, Erwin (2006): *A Frequency Dictionary of German. Core vocabulary for learners*. Routledge: London, New York.
11. Haderlein, Veronika (2008): *Das Konzept zentraler Wortschätze. Bestandsaufnahme, theoretisch-methodische Weiterführung und praktische Untersuchung*. Dissertation. München.
12. Herbst, Thomas/Michael Klotz (2003): *Lexikografie: Eine Einführung*. UTB: Stuttgart.
13. Kaeding, Friedrich W. (Hrsg.) (1897): *Häufigkeitwörterbuch der deutschen Sprache*. Selbstverlag: Steglitz bei Berlin
14. Kempcke, Günter (1992): Organisationsprinzipien und Informationsangebote in einem Lernerwörterbuch. In: Brauße, Ursula/Dieter Viehweger (Hrsg.): *Lexikontheorie und Wörterbuch*. Tübingen: Max Niemeyer, 161–243.

15. Lange, Willi u.a. (2015): Grundwortschatz Deutsch als Fremdsprache: Ein datengeleiteter Ansatz. In: Kilian, Jörg /Jan Eckhoff (Hrsg.): *Deutscher Wortschatz – beschreiben, lernen, lehren. Beiträge zur Wortschatzarbeit in Wissenschaft, Sprachunterricht, Gesellschaft*. Frankfurt am Main u.a.: P. Lang.
16. Lübke, Diethard (2008): *Lernwortschatz Deutsch. Deutsch-Englisch*. Hueber: Ismaning.
17. META-NET (2012): *Europe's Languages in the Digital Age*: (<http://www.meta-net.eu/> Abruf am 3.10.2015)
18. Möhring, Jupp (2011): Möhring, Jupp (2011). Kollokationen im Lernerwörterbuch – Anspruch und Wirklichkeit. *Linguistik online* 47.3: 33–53.
19. Müller-Spitzer, Carolin (2003): Ord nende Betrachtungen zu elektronischen Wörterbüchern und lexikografischen Prozessen. *Lexikographica* 19: 140–168.
20. Ortman n, Wolf D. (1975): *Hochfrequente Wortformen I*. Goethe-Institut: München.
21. Petkova-Kessanlis, Mikaela (2010): Fachsprachliche Kollokationen im linguistischen Diskurs. In: Durčo, Peter (Hrsg.): *Feste Wortverbindungen und Lexikographie. Kolloquium zur Lexikographie und Wörterbuchforschung*. De Gruyter: Berlin, New York, 115-126.

22. Okamura, Saburo /Willi Lange/ Joachim Scharloth (2012): Methoden der Bestimmung des Kernwortschatzes Deutsch. In: Dies. (Hrsg.): *Grundwortschatz Deutsch: Lexikografische und fremdsprachendidaktische Perspektiven*. JGG: Tokyo. 29-44. (= Studienreihe der Japanischen Gesellschaft für Germanistik 088)
23. Schnörch, Ulrich (2002): *Der zentrale Wortschatz des Deutschen. Strategien zu seiner Ermittlung, Analyse und lexikografischen Aufarbeitung*. Narr: Tübingen.
24. Tschirner, Erwin (2005): Korpora, Häufigkeitslisten, Wortschatzerwerb. In: Heine, Antje, Mathilde Hennig, Erwin Tschirner (Hrsg.): *Deutsch als Fremdsprache – Konturen und Perspektiven eines Faches*. München: Iudicium, 133–149.
25. Tschirner, Erwin (2008). *Grund- und Aufbauwortschatz Deutsch als Fremdsprache nach Themen*. Berlin: Cornelsen Verlag.