Erläuterungen zu den Materialdaten

Nachfolgender Text ist gekürzt meiner Dissertation entnommen . Die ausführlichen Erläuterungen zum hier zusammengestellten Material finden sich in: Ulrich Schnörch (2002): Der zentrale Wortschatz des Deutschen. Strategien zu seiner Ermittlung, Analyse und Darstellung (= Studien zur deutschen Sprache). Tübingen.

Die sieben ausgewählten Korpora der Materialgrundlage werden als Untersuchungsgegenstand bezeichnet, was auf zweifache Weise zu verstehen ist: Sie sind einmal Ergebnisse der (Grund-) Wortschatzforschung, zum anderen stellen die einzelnen Korpora das Wortmaterial zur Verfügung, das (vergleichend) analysiert werden kann. Die wichtigsten Eigenschaften der sieben Korpora lassen sich in einer tabellarischen Übersicht zusammenfassen:

	Intendierter Gerbrauchszweck	**Art und Umfang des zugrundegelegten Materials: Erhebungsmethode**
Kosaras: GWS	Lernerwörterbuch (Zweitspracherwerb	__
Pfeffer: Dictionary	Grundlage für die didaktische Anwendung im DaF-Unterricht (Zweitspracherwerb)	Gesprochen-überregional.: 595.000 Wörter aus Tonbandaufnahmen themenbezogener Gespräche im gesamtdeutschen Sprachraum, die auf 25.000 Stichwörter reduziert werden können; davon gehen die 737 häufigsten in die Grundstufe ein. 83.300 Wörter aus themenbezogenen Assoziationstests, die auf 34.000 Stichwörter reduziert werden können; davon gehen 357 in die Grundstufe ein. Ergänzung der Systemlücken: 185 Wörter. Schriftsprachkorpus (für die Mittelstufe): 500.000 Wörter; weitere Aufteilung unbekannt.
Ruoff: HGS	Strukturelle Untersuchung frequenzabhängiger Aspekte des Sprachsystems in gesprochener Sprache (Alltagssprache d. Region)	Gesprochen-regional: 500.000 Wörter aus Tonbandaufnahmen 'freier Gespräche' im Raum Baden-Württemberg; davon entfallen 185.600 auf das 'Schwarzwaldkorpus'. Insgesamt lassen sich 15.676 Lemmata ansetzen
Rosengren: FdZ	Strukturelle Untersuchung frequenzabhängiger Aspekte des Sprachsystems in geschriebener Sprache (Zeitungssprache)	Geschrieben-überregional: ca. 3 Mio. Wörter aus den beiden Zeitungen 'Die Welt' bzw. 'Süddeutsche Zeitung'
Augst: KW	Systemhafte Analyse des Erstspracherwerbs; Vergleich mit dem Worschatz Erwachsener	Gesprochen-kindlich: ca. 38.000 Wörter aus Tonbandaufnahmen von 10 Vorschulkindern über einen Zeitraum von jeweils vier Monaten. Das ergibt ca. 16.000 Stichwörter.
Zertifikat DaF (^{5}1992)	Wort- bzw. Sememminimum für die Erlangung des Zertifikats Deutsch als Fremdsprache: Stoffsammlung, die eine weitergehende didaktische Aufbereitung verlangt.	Gesprochene 'Standardsprache' Die Auswahl erfolgt nach kommunikativ-pragmatischen Grundsätzen durch ein Expertengremium: 2136 Lemmata, die durch 3781 Beispielsätzen illustriert werden (hinzu kommen eine Wortgruppenliste sowie eine Wortbildungsliste, welche direkt oder indirekt den tatsächlichen Bestand an Lemmata erhöhen).
Krohn: KLK	Synthetische Wortliste: Schnittmenge* zur Untersuchung des themenunspezifischen Wortschatzes *(Kontaktschwelle, Zertifikat, Käding/Meier, Pfeffer I & III, Rosengren, Erk)	Schnittmengenbildung 1352 Wörter: Autosemantika (ohne Synsemantika, ca. 100 ohne nähere Angabe)

	Aufbereitung: **Berücksichtigung der Semantik**	**Darbietung** **des Materials**	**Zähleinheiten im PC-Korpus**
Kosaras: GWS	ja, durch Interpretamente Beispielsätze (sowie durch Wortbildungen und Kollokationen)	Nach semasiologischen Prinzipien aufgebautes, alphabetisch geordnetes, einsprachiges Wörterbuch	4976 Sememe
Pfeffer: Dictionary	Zähleinheit ist primär das Wort, Bedeutungen werden aber unterschieden: (teilweise) frei vs. (teilweise) begrenzt. Bedeutungserklärung durch Beispielsätze und englische Übersetzungsäqvivalente (auch Angabe von Idioms)	Nach semasiologischen Prinzipien aufgebautes, alphabetisch geordnetes zweisprachiges Wörterbuch	4943 Sememe
Ruoff: HGS	Zähleinheit ist das Wort. Im Vertrauen auf die Kompetenz des 'native speakers' wird i.d.R. keine Bedeutungsdifferenzierung vorgenommen (Ausnahme: Unterscheidung von Homographen). In Zweifelsfällen: zusätzliche Bedeutungsangabe	Drei nach den Kriterien des Alphabets, der Wortart und der Frequenz geordnete Listen	4023 Wörter
Rosengren: FdZ	nein Lemmatisierung nach dem Kriterium der 'Flexionsreihe' – auf dieser Basis auch Unterscheidung von homonymen und polysemen Wörtern	Wortformen (FdZ1) bzw. Lemmata (FdZ2) listenförmig, nach verschiedenen Kriterien sortiert	4050 Wörter
Augst: KW	Zähleinheit ist primär das Wort: Bedeutungen werden unterschieden durch die Zuordnung des jeweiligen Lemmas in eine oder mehrere Kleinstfelder. Das ergibt 23.403 Einträge (Sememe?). Keine Unterscheidung zwischen Homonymie und Polysemie. In Zweifelsfällen: Zusätzliche Bedeutungsangabe	Nach onomasiologischen Kriterien aufgebautes Wörterbuch (Übernahme der Begriffssystematik von Wehrle/Eggers)	3722 (Stich-)Wörter
Zertifikat DaF (^{5}1992)	ja, durch Beispielsätze	Wortliste mit Beispielsätzen	3781 Sememe
Krohn: KLK	nein	Wortliste	1352 Wörter:

Zusammenfassung der wichtigsten Charakteristika des Untersuchungsgegenstandes

Auf der Grundlage dieser sieben Korpora sind nun die genaueren Vorbedingungen des Wortschatzvergleichs, also der Schnittmengenbildung und der anschließenden Wortschatzanalyse zu klären. Die erste Voraussetzung für eine Wortschatzuntersuchung ist die Erfassung des jeweiligen Sprachdatenmaterials mittels EDV. Am Beginn jeder Daten-Verarbeitung steht folglich die Aufbereitung der zugrunde liegenden Wortschatzsammlung, d.h. die sieben Korpora mussten zunächst in einer Datenbank erfasst werden.

Das auf dieser Stufe vorrangig angestrebte Ziel war es, die (zentralen) Wortschätze der einzelnen Korpora vergleichen, also v.a. rechnersortierte Schnittmengen bilden zu können. Mit dieser Vorgabe ging es als Erstes darum, die sieben einzelnen Computerlisten nacheinander zu erstellen. Die wesentlichen Punkte dabei sind:

- Jede der Listen soll für sich einen charakteristischen Extrakt der ihr zugrundeliegenden Quelle repräsentieren; was zur Folge hat, dass über den reinen Wortbestand hinaus die entsprechenden Zusatzinformationen einzubeziehen sind, wobei der Grundsatz gilt, bereits bei der Eingabe lieber mehr als zu wenig Informationen zu berücksichtigen, um den Spielraum für die späteren Analysen von vorherein möglichst groß zu halten.

- Der Informationsgehalt der jeweiligen Quellen ist aber gemäß ihres Gebrauchszwecks mitunter sehr verschieden(artig). Damit man alle relevanten Angaben erfassen kann, sind jeweils individuell zu wählende Datenbankstrukturen nötig, die sich zwangsläufig bis zu einem gewissen Grad unterscheiden.

- Da die sieben Listen schließlich in eine zusammenführt werden sollen, muß jedoch ein Grundbestand an Gemeinsamkeiten erhalten bleiben. Die jeweiligen Datenbankstrukturen müssen daher insgesamt vereinfacht bzw. vereinheitlicht werden, wobei natürlich Schwerpunkte zu setzen sind, um einerseits die Vergleichbarkeit selbst zu gewährleisten, um andererseits für den Vergleich wichtige Informationen von vornherein berücksichtigen zu können.

Es ist klar, dass auf Grund dieser Vorüberlegungen jede Einzeldatenbank in gewisser Weise eine Kompromisslösung darstellt.

Jene sieben auf die beschriebene Weise erstellten Datenbankfassungen der Korpora stellen fortan die Grundlage für alle weiteren Analyseschritte dar. Abgesehen von Formalia abgesehen lässt sich die Mehrzahl der zu lösenden Detailprobleme i.d.R. folgenden Schwerpunkten zuordnen: die Vergleichsgrundlage als solche (quantitative Ausgewogenheit), die Lemmatisierung (Definition der Zähleinheit) und Bedeutungsangabe(n) sowie die Wortartenzuordnung.

Die sieben Computerlisten bilden die Brücke zur eigentlichen Analyse. Bei der Digitalisierung der Korpora wurde darauf geachtet, dass stets drei der Felder bei allen Korpora identisch sind: Wort, Korpus und Wortart, hinzu kommen die entsprechenden Spalten, welche Informationen zur Bedeutungsunterscheidung bei Kosaras und Pfeffer enthalten. Damit ist eine Vorbedingung für eine spätere Schnittmengenbildung erfüllt, v.a. aber für die semantische Analyse einer Schnittmenge, die deshalb in jedem Fall mit Kosaras und Pfeffer gebildet werden muss.

Die Voraussetzung für die nächsten Arbeitsschritte ist, die sieben Einzeldatenbanken in eine Gesamtliste zusammenzuführen. Die Einzeldateien wurden also zunächst einfach addiert und anschließend nach der Spalte ‘Wort’ fortlaufend alphabetisch sortiert. Das Ergebnis ist eine umfasssende Computerdatenbank, welche die Informationen aller sieben Korpora in sich vereint. Sie umfaßt insgesamt 26847 Datensätze (= Summe der Datensätze aller Einzeldateien), d.h. ebensoviele nach Form, ggf. Inhalt und nicht zuletzt nach Quell-Korpus unterschiedener Zähleinheiten.

Für die Gesamtliste bedeutet dieses Vorgehen nun dreierlei:

- Auf engem Raum wird übersichtlich die Information aus **allen** Dateien aufgelistet bzw. einander gegenübergestellt.
- Während die Spalten Wort, Korpus und Wortart dateienunabhängig sind, beinhalten alle übrigen dateien-, d.h. korpusspezifische Informationen, die ggf. den Spielraum für weiterreichende Interpretationen erhöhen können.
- Die Schnittmengen, aber auch die Mengen der Hapaxlegomena können auf der Basis dieser Gesamtliste ermittelt werden.

Die kombinierte Gesamtliste vereint also die Summe aller Einzelschritte der Datenerfassung und stellt zugleich das erste Ergebnis der Datenverarbeitung bzw. –analyse dar. Sie bildet aber gleichermaßen die Grundlage für die Untersuchung und (vergleichende) Auswertung der Korpora, die im Textband in zwei Schritten erfolgt. Diese Materialbasis umfasst im Wesentlichen fünf Beleglisten, die aus der Gesamtliste generiert wurden und als PDF-Files abgerufen werden können.

Alle Ergebnisse der Detailanalyse habe ich in die (entsprechend vorbereitete) Datenbank der Untersuchungsschnittmenge eingetragen. Für die Präsentation der drei wortartenspezifischen Teilmengen wurde die ursprüngliche Tabellenform aufgegeben zugunsten einer an traditionellen Wortschatz-Karteikarten orientierten, übersichlicheren Darstellung mit folgender Struktur:

[Anzahl] **Lesart**	Frequenzen: 'Dictionary': FdZ: NR_W, - NR_{SZ}, - AF_W, - AF_{SZ}, - Ü; relativ;	HGS: absolut – relativ;
Morphologie + Grammatik (bei V, S)/Wortart (bei Sonstigen) + Morphologie bei Adj.		
Worterklärung (nach Kosaras 1980 = K)		
Beispielsatz/-sätze (nach Pfeffer (1970 = p)		
Synonyme (nach Görner/Kempcke [7]1988)		
Antonyme (nach Agricola, Chr./Agricola, E. [2]1992)		
Diverses: Bemerkungen/Nachträge u.ä.		

Das Gesamtmaterial ist in Form zweier Lemmalisten initialalaphabetisch bzw. rückläufig alphabetisch sortiert, wobei zusätzlich angegeben wird, in welchem Korpus bzw. in welchen Korpora die einzelnen Lemmata vorkommen.