Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 1, Nummer 2 (Dezember 2021), ISSN 2749-4411

KORPUSFREQUENZEN UND ANDERE METRIKEN ZUR STRUKTURIERUNG VON DAF-LEHRMATERIAL

Rainer Perkuhn
Leibniz-Institut für Deutsche Sprache (Mannheim)

Abstract

Korpora und Fremdsprachendidaktik haben – auch jenseits des angeleiteten oder selbstgesteuerten Arbeitens an den Daten – Berührungspunkte mit langer Tradition, durchaus mit nicht-digitalen Ausläufern, deren korpuslinguistische Dimensionen erst in den letzten Jahrzehnten erschlossen wurden. Worthäufigkeitszählungen, auch vergleichend, in beliebig großen oder auf bestimmte Bedürfnisse zugeschnittenen Datensammlungen lassen sich mit weiteren Metriken verknüpfen, die eine differenzierte Bewertung für die didaktische Relevanz ermöglichen. Kollokations-/Kookkurrenzanalysen helfen, typische Formulierungsmuster zu ermitteln. Dieser Beitrag stellt zunächst diese beiden Herangehensweisen dar. Das Manko der getrennten Betrachtung ist, dass keine der beiden isoliert ausreicht, um die Angemessenheit von Formulierungen zu bewerten hinsichtlich muttersprachlicher Natürlichkeit *und* Weiterentwicklung des Lernstands. Als Abhilfe wird eine Verknüpfung skizziert, die beide Perspektiven zusammenbringt.

Keywords: Korpuslinguistik; DaF; Wortschätze; Syntagmatik; Wortfrequenzen; Streuung / Dispersion; Kollokation / Kookkurrenz

Abstract

Corpora and foreign language didactics have – even beyond guided or self-directed practice with the data – points of contact with a long tradition, with non-digital spin-offs, whose corpus linguistic dimensions emerged in the last decades. Frequency counts, also comparative, in data collections of any size or tailored to specific needs combined with further metrics allow a differentiated didactic assessment. Collocation analyses identify typical formulation patterns. This paper first presents these two approaches. The shortcoming of considering them separately is that neither of them in isolation evaluates adequately the appropriateness of formulations with respect to native-like naturalness and learning progress. As a remedy, a linkage is outlined that brings both perspectives together.

Keywords: Corpus Linguistics; GFL; Vocabularies; Syntagmatics; Word Frequencies; Mean Variation / Dispersion; Collocation / Cooccurrence

Alles, was man wissen muss, um eine Sprache zu erwerben/erlernen/vermitteln, steckt in der Sprache selbst.
(nach Perkuhn / Belica 2006: 7)

1. Einleitung[1]

So programmatisch schön das oben genannte Zitat als Leitthese klingt, so sehr leidet diese doch unter einem Haken: Für den ‚unbewussten' Prozess des Erwerbs der Muttersprache ergibt sich die Auseinandersetzung mit der ‚Sprache selbst' auf eine natürliche Art und Weise. Für das angeleitete ‚bewusste' Sprachenlernen/-lehren können wir die ‚Sprache selbst' als Objekt aber nicht so fassen,

[1] Der Aufbau und alle wesentlichen Punkte der Argumentation wurden in einem Vortrag auf dem GAL2020 Symposium ‚Korpora – eine Chance für DaM/DaF/DaZ: Theorie und Praxis' am 10.9.2020 vorgestellt, die Einleitung entspricht weitestgehend dem dort eingereichten Abstract.

dass wir die nötigen Informationen kondensieren können, schon gar nicht in ‚mundgerechten Portionen'. Als Ersatzobjekt bedient man sich in der Korpuslinguistik digitaler Textsammlungen, die nach unterschiedlichen Gesichtspunkten zusammengestellt und aus verschiedenen Perspektiven ausgewertet werden können. Die Beschäftigung mit diversen Analyseverfahren und Metriken mit entsprechendem Blick auf ausgewählte Daten zeigt das Potential für die Anwendung in der Fremdsprachendidaktik, da es im Folgenden um deutschsprachige Textsammlungen geht, entsprechend für Deutsch als Fremdsprache (DaF).

1.1 (Digitale) Textsammlungen für den Unterricht

Jede Textsammlung kann ein Gewinn für den Fremdsprachenunterricht sein, sei es als Lektüre oder als Quelle der Inspiration für Lehrende und Lernende. Sofern die Textsammlungen digital als sogenannte Korpora vorliegen, erweitert sich das Spektrum diverser Nutzungsszenarien. Dazu gehören z.B. diverse Möglichkeiten unterschiedlicher Arten der Recherche, auch zu Wortbildungen, sowie die Auswertung von ggf. gefilterten und sortierten Übersichten, Konkordanzen und Ansichten längerer Textabschnitte (vgl. Sinclair 2004; McEnery / Xiao 2010). Ein naheliegendes Szenario ist die Suche über, aber auch nach sprachlichen Einheiten unterschiedlicher Granularität mit bestimmten Eigenschaften: z.B. Wörter, Formulierungen, Sätze oder auch Texte, die einem gegebenen oder angestrebten Lernerstand entsprechen (vgl. Fandrych / Tschirner 2007; Ahrenholz / Wallner 2013; Wallner 2013). Auch Texte von Lernenden können als sogenanntes Lernerkorpus erfasst werden, um Stand und idealerweise sogar Lernfortschritt auswerten zu können (vgl. Granger 2017; Lüdeling / Walter 2009, 2010).

Mithilfe des Computers lassen sich die verwendeten Wörter und Formulierungen einer Textsammlung zu denen einer anderen oder zu einer vorgegebenen Wortliste abgleichen, sodass sie in Bezug auf Lernstand oder Referenzniveau eingestuft werden können. Je größer die Korpora werden, desto wichtiger wird es, dass für den Computer Operationalisierungen für linguistische Begriffe zur Verfügung stehen, was allerdings mit steigender Abstraktionsstufe, bereits beginnend mit dem Begriff ‚Wort', nur mit abnehmender Zuverlässigkeit gegeben ist. Mit diesem – zugegebenermaßen teilweise pragmatischen – Vorbehalt und der offen gehaltenen Auswahl eines Korpus, kann der Computer seine Stärken bei der Anwendung quantitativer Verfahren einbringen: Häufigkeiten der sprachlichen Einheiten (zunächst meist Wörter oder Wortformen) zählen und deren Vorkommen im Kontext anderer Wörter oder in Textmengen bestimmter Eigenschaften (statistisch) bewerten. Dies führt zu Frequenzlisten, Streuungsmaßen (sog. Dispersion) oder auch zu thematischen (o.ä.) Schlüsselworteigenschaften (sog. *Keyness*). Bereits die Beschäftigung mit dem ersten Aspekt hat für die Fremdsprachendidaktik eine große Wirkung entfaltet. Mit Bezug auf die Arbeit von Nation (insbesondere Nation 2001) wird oft die Wichtigkeit der häufigsten Wörter herausgestellt. Diese sollten möglichst früh vermittelt werden, da z.B. mit den 2.000 häufigsten Wörtern 80% des Vokabulars durchschnittlicher (Zeitungs-)Texte abdeckt werden. Genau genommen hat Nation allerdings 2.000 Wortfamilien angesetzt, basierend auf einem gewissermaßen erweiterten Grundform- oder Stichwortbegriff. Nation (2016) diskutiert ausführlich die Herausforderungen bei der Umsetzung der Begriffe wie auch des gesamten Vorgehens bei der Ermittlung von Wortlisten, die stets nur einen annähernden Charakter haben können.

So beeindruckend der Wert 80% sein mag – der sich im Übrigen für keine sprachliche Einheit an deutschen Korpora bestätigen lässt (vgl. Abb. 1) - so sollte auch evident sein, dass dem/der Lernenden durchschnittlich jedes fünfte Wort unbekannt ist. Kyongho / Nation (1989) zitieren in diesem Sinne Laufer (1986), nach dem 95% Textüberdeckung für das Verständnis notwendig sei. Der Kerngedanke, um die Lücke zu schließen, kann demselben Aufsatz entnommen werden in Kombination mit Clarke / Nation (1980) oder auch Mukherjee (2002): Die Lernenden sollen versuchen, sich die unbekannten Wörter selber aus dem Kontext zu erschließen. An dieser Stelle kann ebenfalls korpuslinguistische Methodik unterstützen, indem das Musterhafte der Kontexte aufgespürt wird. Ergänzend zu den Ideen von Hausmann (1984, 2004) hat sich ein eher weiter Kollokations-/Kookkurrenzbegriff etabliert, wie ihn Bahns (1997) oder auch Belica / Perkuhn (2015) – Sinclair (1991) folgend – vertreten. Ein aussichtsreicher Ansatz, der hier skizziert werden soll, wäre von einem nach Häufigkeit ausgewählten (oder angepassten) Kernwortschatz auszugehen, der systematisch zu themenbezogenen Wortverbindungen erweitert wird, und dann möglichst zu abgeschlossenen Zuständen gestufter einfacher Sprache im Sinne von Baumert (2016) konvergiert – ganz im Sinne didaktisch-methodischer Prinzipien: Vom Leichten zum Schwierigen, vom Einfachen zum Komplexen, vom Bekannten zum Unbekannten.

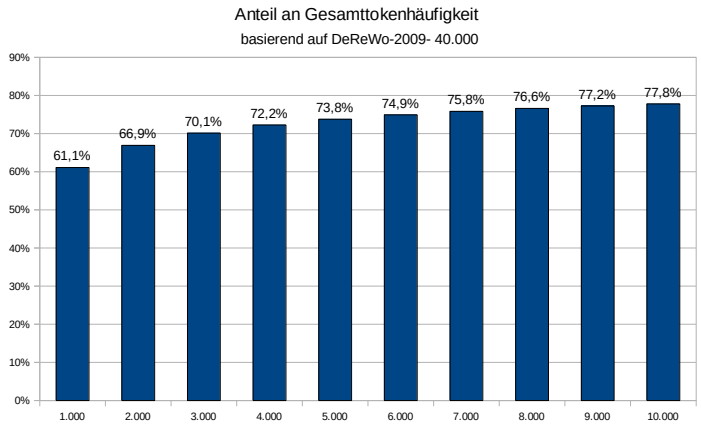


Abb. 1
Textabdeckung der ersten N tausend Lemmata (N = 1 .. 10)[2]

1.2 Setting

Der Erwerb der Muttersprache gliedert sich auf natürliche Art und Weise in mehrere Abschnitte, von denen allerdings nur die allerersten bis zum Grundschulalter gut beschrieben sind (vgl. Kegel 1974; Klann-Delius 1999). Nicht nur deshalb lässt sich dieser Prozess nicht 1:1 auf das Vermitteln bzw. Lernen einer Fremdsprache übertragen. Trotzdem muss das Lernen auch in (evtl. grobe) Etappen und für Unterrichtseinheiten fein unterteilt werden bis hin zur Auswahl des sprachlichen Phänomens, das als nächstes angegangen werden soll.

[2] Basierend auf http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-40000g-2009-12-31-0.1.zip (05.11.2021).

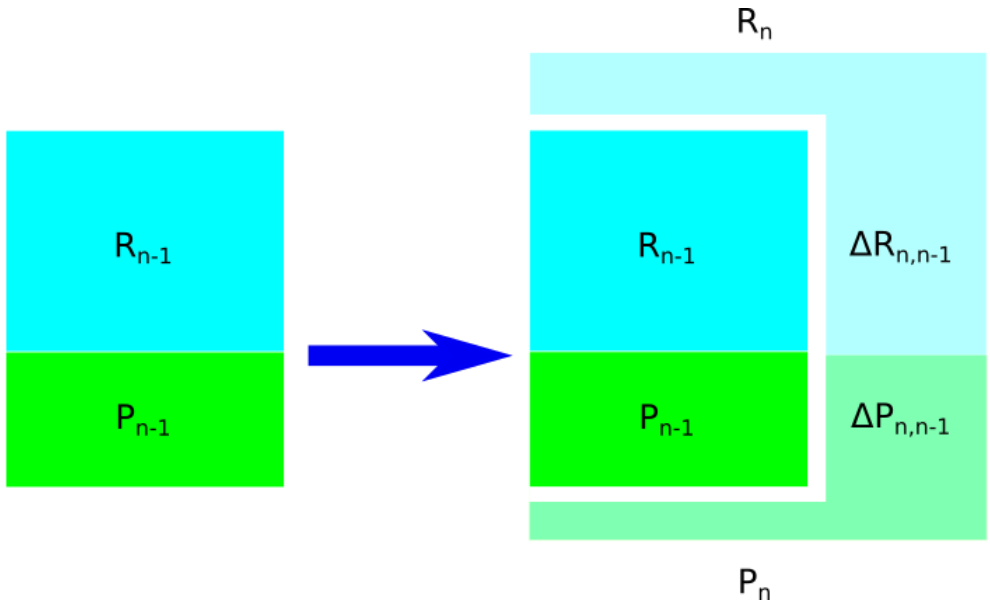


Abb. 2
Modell des Lernfortschritts von Stand n-1 zu Stand n (‚eine Lernetappe')

Mit einem abstrakten Modell soll der Übergang von einem Zustand zum nächsten unabhängig von der tatsächlichen Granularität veranschaulicht werden (vgl. Abb. 2): Ausgehend von einem Vorzustand n-1 sollen sowohl die rezeptiven (R) als auch die produktiven Fähigkeiten (P) weiter entwickelt werden. Das Modell nimmt vereinfachend Monotonie an, d.h., dass keine Fähigkeiten verloren gehen oder sich wandeln (z.B. von R nach P). Neben dem Aspekt des beschreibenden IST-Zustandes kann der Folgezustand auch als Planungsziel ausgelegt werden. Besondere Zustände wären etwa R_0 als Ausgangszustand mit keinen Kenntnissen und R_{fin} als Zustand des angestrebten Lernendniveaus; für die aufeinander aufbauenden Niveaus des Gemeinsamen Europäischen Referenzrahmens (vgl. Europarat 2001) ließen sich z.B. R_{B1} oder R_{B2} formulieren, sowie $\Delta R_{B2,B1}$ als Zugewinn einer Lernetappe.

Aus didaktischer Sicht wäre von den oben genannten Sortier- und Filterkriterien besonders die Möglichkeit interessant, Texte (oder auch Teile davon) einem Lernstand und ggf. auch einem Lernfortschritt zuordnen zu können. Dies erfordert eine Bewertungsfunktion, die – gegeben einen oder zwei Lernstände – angibt, zu welchem Grad eine sprachliche Einheit diesen Ständen entspricht. Für Testmaterial z.B. zum Niveau R_{B2} sollten die sprachlichen Einheiten zu (nahezu) 100% in R_{B2}, das Abgeprüfte speziell in P_{B2}, eingeordnet werden können, für Lern-/Übungsmaterial zu $\Delta R_{B2,B1}$ sollten die sprachlichen Einheiten zu einem hohen Anteil zu $R_{B1,}$ zu einem kleinen Anteil zu R_{B2} bzw. P_{B2} passen – die genauen Werte variieren mit der Intensität des Lernkurses. Liegt die Bewertungsfunktion vor, so lässt sie sich nicht nur in die oben zuerst genannte Korpusarbeit integrieren. Sie kann für jede Sammlung von sprachlichen Einheiten, auch von ausgewählten, angepassten oder vollkommen künstlichen Beispielen, dazu dienen, diese als Lehrmaterial (im weitesten Sinne) zu strukturieren, um das Sortieren und Filtern der Sammlung überhaupt erst zu ermöglichen. Idealerweise wird dadurch Sprachinput analog zum Erstspracherwerb, allerdings in kondensierter, gestaffelter Form angeboten werden können.

Bisherige korpusbasierte Ansätze der „Bewertungsfunktion" konzentrieren sich dabei weitestgehend auf redaktionell vorgegebene (z.B. vom Goethe-Institut), ggf. zusätzlich thematisch

gegliederte Wortschätze[3], sowie einfache Metriken wie Satzlänge oder Anteil Funktionswörter pro Satz. Im Folgenden werden die verschiedenen Dimensionen der Bewertungen vorgestellt, diejenigen, die bereits an anderen Stellen zusammengefasst wurden, nur kurz. Etwas ausführlicher dargestellt werden ein Aspekt, dem bisher erst wenig Aufmerksamkeit zuteilwurde, sowie das eigentliche Anliegen dieses Beitrags, das Zusammenbringen der Dimensionen zu skizzieren.

2. Wortschatz und Frequenz

Einen Teil des Lernstoffs macht im herkömmlichen Sinne Wortschatzarbeit aus, d.h. eine Liste der nächsten zu lernenden Vokabeln als Teil des gerade genannten geplanten Zugewinns. Um herauszufinden, nach welchen Kriterien diese Teilwortschätze z.B. für Lehrwerke ausgewählt werden, haben wir den Themenbereich ‚Wohnen' der Lehrmaterialsammlung *Profile Deutsch* (vgl. Glaboniat / Müller / Rusch 2002) ausgewertet und mit Häufigkeiten einer korpusbasierten Grundformfrequenzliste abgeglichen (vgl. Abb. 3). Immerhin wird auch bei der Beschreibung der Niveaukompetenzen an verschiedenen Stellen auf Häufigkeit (oder Gebräuchlichkeit) verwiesen, für „A2– Grundlegende Kenntnisse" etwa in der Form: „Kann Sätze und häufig gebrauchte Ausdrücke verstehen"[4].

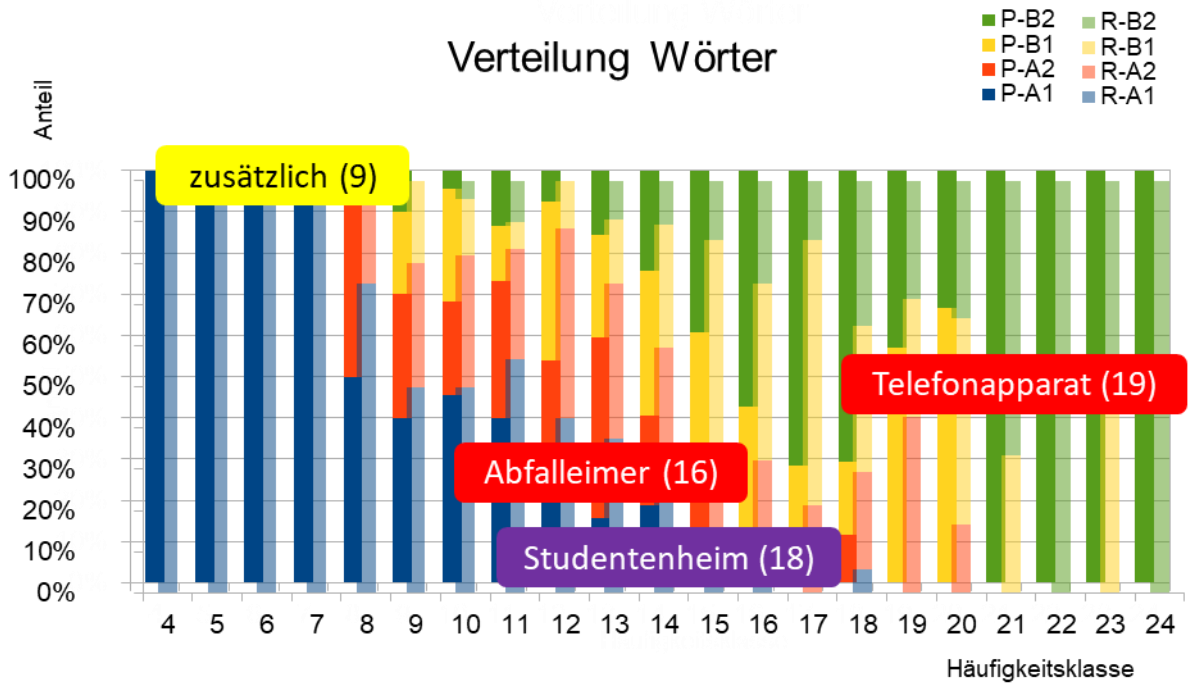


Abb. 3
Abgleich Themenwortschätze ‚Wohnen' aus *Profile Deutsch* mit korpusbasierter Häufigkeitsliste DeReWo[5]

[3] Siehe aktuelle Arbeiten zu Korpora gesprochener Sprache, z.B. des Projekts ZuMult (https://zumult.org/) (05.11.2021).
[4] https://www.goethe.de/z/50/commeuro/303.htm (05.11.2021).
[5] http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-40000g-2009-12-31-0.1.zip (05.11.2021).

Beim Vergleich der Teilwortschätze der verschiedenen Niveaustufen produktiv und rezeptiv von A1 bis B2 lässt sich der Trend feststellen, dass die häufigeren Wörter in den frühen Phasen, seltenere Wörter in den späteren Phasen geplant sind. Dass ein (relativ) häufiges Wort eventuell später als erwartet berücksichtigt wird, mag daran liegen, dass es ein etwas anspruchsvolleres Funktionswort ist (z.B. *zusätzlich*). Der Umstand, dass Wörter, die relativ früh vermittelt werden sollen, in dem Korpus nicht so häufig belegt sind, kann auf unterschiedliche Art gedeutet werden: Entweder ist der gesetzte thematische Schwerpunkt (oder das Wortfeld) nicht ausreichend in dem Korpus berücksichtigt – oder es handelt sich um ‚Altlasten', Wörter, die früher für wichtig gehalten, inzwischen aber von anderen Ausdrücken abgelöst wurden. Betrachtet man die in Abb. 3 hervorgehobenen Beispiele, stellt man fest, dass es modernere und gebräuchlichere alternative Ausdrücke gibt (vgl. Abb. 4).

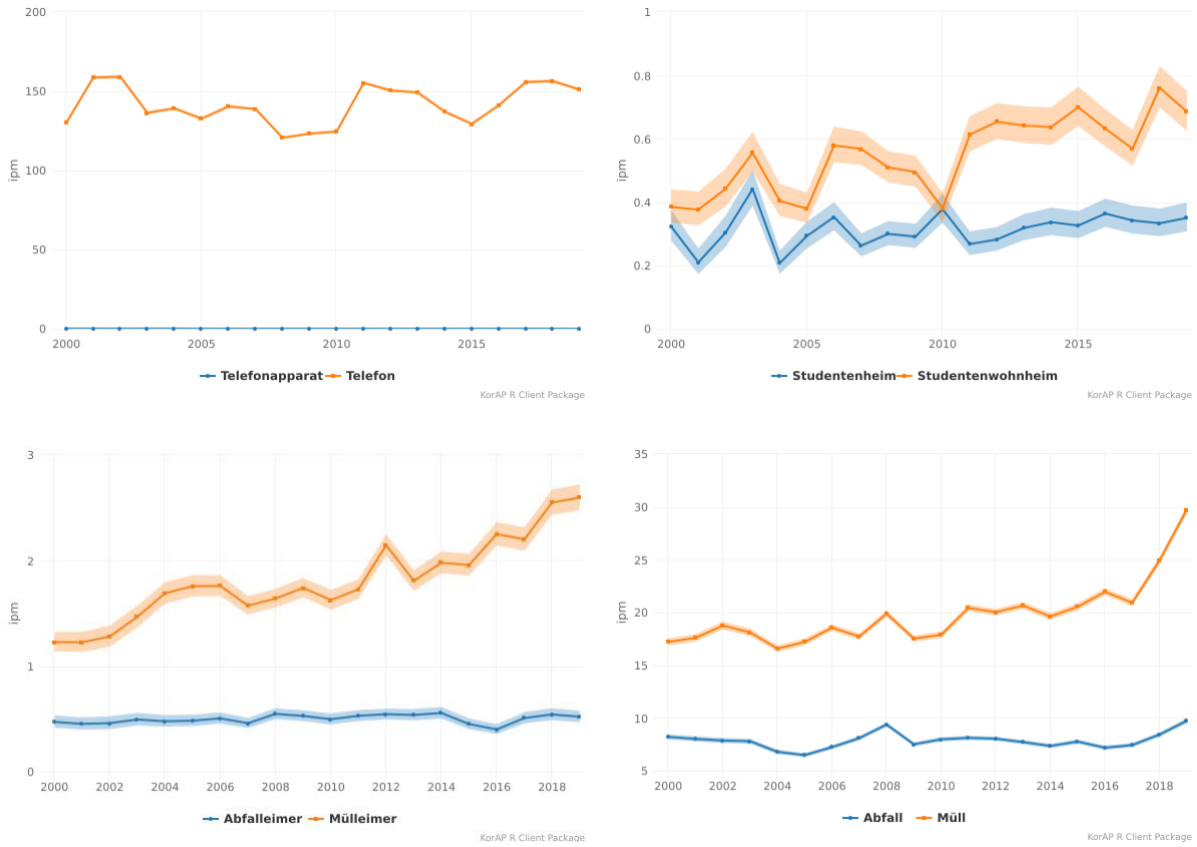


Abb. 4
Zeitverläufe zu den drei rechten Beispielen aus Abb. 3 sowie zum Erstglied des dritten Beispiels (erzeugt mit RKorAP-Client auf DeReKo-2020-I)[6]

Tatsächlich wirkt bei vielen Wortschatzlisten nach, dass sie in langer Tradition von älteren Häufigkeitslisten abgeleitet wurden, beginnend bei Kaeding (1897/98) und später dem *Zertifikatswortschatz Deutsch*, wie auch bereits Tschirner (2005: 136) festgehalten hat. Mittlerweile greifen aber auch die Herausgeber von Lehrmaterial auf aktuellere Häufigkeitslisten zurück. Im

[6] Die Konfidenzintervalle sind zum Teil kaum zu erkennen, da die Häufigkeiten um Größenordnungen auseinander liegen, *Telefonapparat* liegt etwa absolut im kleinen zweistelligen Bereich, ipm < 0,1, ipm = relative Häufigkeit als „Instanzen pro Million".

Vorwort zum Langenscheidt *Großwörterbuch Deutsch als Fremdsprache* ist zum Punkt „1.2 Der Zentralwortschatz“ vermerkt:

> In diesem Wörterbuch sind ca. 5500 Stichwörter mit dem Zeichen ★ markiert. Dabei handelt es sich um Wörter, die in verschiedenen sogenannten Grundwortschätzen und Aufbauwortschätzen enthalten sind und für viele Sprachprüfungen benötigt werden. Ergänzt wurden diese Listen um Wörter, die in den Häufigkeitslisten (DeReWo) des Instituts für deutsche Sprache (IDS) und des Projekts Deutscher Wortschatz der Universität Leipzig zu den 4000 häufigsten Wörtern gehören, soweit sie für Deutschlernende wichtig erschienen. (Götz 2019: 7) [Unterstreichung vom Autor, s. Fußnote][7]

Sofern die im Korpus abgebildete Sprache ungefähr den Zielvorstellungen entspricht, auf die ein Sprachlernkurs ausgerichtet ist, ist es plausibel, für das Fernziel die Häufigkeiten im Gesamtbestand zu Grunde zu legen. Wenn der Kurs allerdings auf kleinere Einheiten heruntergebrochen wird, stellt sich die Frage, ob der gleiche Gedanke dahingehend übertragen werden kann. Die allerhäufigsten Wörter sind in der Regel tatsächlich wichtige Funktionswörter; es erscheint für ein sinnvoll aufeinander aufbauendes Lernen (sozusagen ein *Bootstrapping*) allerdings wenig sinnvoll, diese alle auf einmal zu vermitteln, genauso wenig, wie zu erwarten ist, dass alle zu einem Thema passenden Wörter jeweils in den folgenden Abschnitten einer Häufigkeitsliste enthalten sind.

Aus der Anwendungsperspektive heraus wird hier und im Folgenden oft ein intuitiver Wortbegriff angesetzt, der einem lexikographischen Lemma nahekommt, d.h. die Gesamtheit eines Flexionsparadigmas mit einer ausgewählten Form benannt, eventuell bereits untergliedert nach Lesarten. Für die empirische Operationalisierung müsste dieser Begriff jeweils sprachspezifisch umgesetzt werden. Eventuell kann auch eine differenzierte Betrachtung je nach Lernstand von unterschiedlichen Wortform- und Lemmabegriffen sinnvoll sein, wie es auch bei Wörterbüchern für den DaM-Primarbereich[8] üblich ist. Eine besondere Beachtung erfordern etwa unregelmäßig gebildete Formen oder besondere Formen in bestimmten Registern, für die es keine einheitliche, übergreifende Lemmazuordnung gibt (wie z.B. *haste* in umschreibender Wiedergabe gesprochener Sprache).

3. Frequenzen und Verteiltheit

Gries / Ellis (2015: 232) stellen Verbindungen her von korpusbasierten statistischen Maßen zu psycholinguistischen und kognitiven Konzepten, die auch für Lerntheorien berücksichtigt werden sollten. Sie machen darauf aufmerksam, dass Frequenzen alleine wenig aussagekräftig sind, wenn nicht berücksichtigt wird, wie sich die Vorkommen in einem Korpus verteilen. Bevor der allgemeine Fall betrachtet wird, soll zunächst der Spezialfall der *Keyness* herausgegriffen werden.

3.1 Keyness

Wenn sich eine Folgeetappe eines Sprachlernkurses auf einen bestimmten Aspekt der Sprache bezieht, sei es insbesondere mündliche oder internetbasierte Kommunikation oder auch einfach ein bestimmtes Themengebiet, lassen sich weitere Maße heranziehen, sofern auch entsprechende Daten,

[7] In älteren Fassungen stand anstelle der unterstrichenen Passage noch: „insbesondere im Zertifikatswortschatz des Goethe-Instituts“.

[8] Der schulische Bereich der 1 bis 4. (teilweise auch bis zur 6.) Klasse für den Unterricht „Deutsch als Muttersprache“.

entweder als eigenes Korpus oder als Ausschnitt eines Referenzkorpus, zur Verfügung stehen. Diese Daten können dann im Vergleich zu den anderen bzw. restlichen Daten, evtl. auch zu Daten des Vorzustands ausgewertet werden. Neben einer reinen Häufigkeitszählung und damit verbunden einem Rangvergleich sind vor allem statistische Bewertungen interessant, die aufdecken helfen, welche Wörter in dem neuen Sprachausschnitt verhältnismäßig überrepräsentiert sind. Sie bewerten die sogenannte *Keyness* dieser Wörter und werden in anderen Zusammenhängen auch für allgemeine Dokumentenklassifikation (wie im *Information Retrieval*, vgl. Sparck Jones 1971) oder konkreter für die thematische Klassifikation von Texten verwendet. Für den DaF-Zusammenhang erhofft man sich dadurch Hinweise auf die Relevanz für einen bestimmten, evtl. auch im weitesten Sinne „thematisch" motivierten Teilwortschatz (vgl. Scott / Tribble 2006).

Tab. 1 zeigt die explorative Anwendung auf zwei sehr große Korpora, um deren *sampling*-bedingte[9] Zusammensetzung anhand weiter herauszuarbeitender Eigenschaften aufzudecken. *Keyness* als Eigenschaft einzelner Terme, gute Prädiktoren für bestimmte Textgruppen zu sein, wird seit einiger Zeit in einem weiteren Sinne verwendet, um evtl. neue Themen und/oder Diskurse (ggf. auch deren Wandel) aufzuspüren. Neben der hierfür zu klärenden *Keyness*-Berechnung selber[10] können dazu selbstverständlich nicht vereinzelte Schlüsselwörter beliebig herausgegriffen werden. Während bei heutigen Ansätzen z.B. Clustering-Verfahren zum Einsatz kommen, wurden die Erkenntnisse in der zitierten Studie durch Auswertungen auf anderen Ebenen, u.a. einer automatischen thematischen Klassifikation, gegengeprüft und abgesichert.

DeReKo-Keywords		**deWaC-Keywords**	
Mark	Franken	ich	Ich
Prozent	Samstag	und	Klägerin
gestern	SPD	…	…
Millionen	…	du	BGB
sei	CDU	Sie	…
Schilling	…	Du	Kläger
sagte	Jahr	mir	…
…	fünf	kann	Gottes
Sonntag	Trainer	mich	…
…	Dollar	Hallo	Beklagte
vergangenen	Polizei	wir	…
Milliarden	Jahren	Gott	Jesus

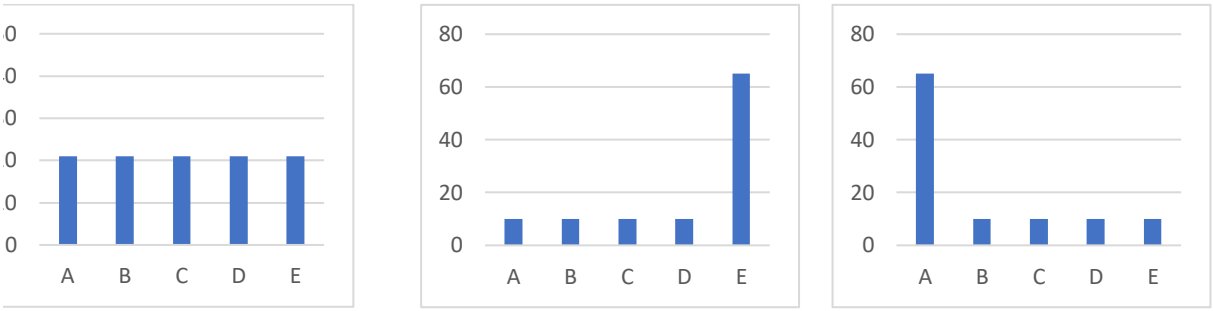
Tab. 1
Aus Darstellungsgründen ausgewählte Top-Keywords einer vergleichenden Analyse DeReKo vs. deWaC (vgl. Belica et al. 2007)

Sowohl thematische als auch Registerbesonderheiten sind gut zu erkennen, z.B. verweisen Mark / Schilling / Franken (und jeweils weitere) auf Währung / Finanzen, SPD / CDU auf Politik, Trainer auf Sport vs. Klägerin / Kläger / Beklagte auf Recht, sowie Gott / Jesus auf Religion; Personalpronomen bzw. bestimmte flektierte Verbformen kennzeichnen unterschiedlichen Stil, berichtend vs. dialogisch. Die Hinweise auf das Register waren nur sichtbar, da der Vergleich auf Wortformebene durchgeführt wurde, ein Vergleich auf Lemmaebene hätte die Unterschiede ausgeblendet gehabt.

[9] D.h. durch Effekte bei der Datenauswahl für die Stichprobe, sowohl konzeptioneller als auch technischer Art.
[10] Siehe dazu (Gabrielatos 2018) zur Diskussion von Assoziationsmaßen, Rangvergleichen, Effektstärken (mit und ohne Signifikanzabsicherung) und Clustering. In der hier zitierten Studie wurde als Assoziationsmaß LLR ($p<0.05$) verwendet, da es vorrangig um die Unterschiede zweier fast gleich großer Korpora ging.

3.2 Streuung / Dispersion

Wenn nicht nur zwei, sondern mehr Datenbestände verglichen werden sollen, interessiert analog die Wohlverteiltheit über die verschiedenen Datenbestände hinweg, d.h. man möchte wissen, ob ein Wort in jedem Teilbestand ungefähr so oft vorkommt, wie es vom Gesamtvorkommen (und somit vom Durchschnitt) her zu erwarten wäre – oder eben sich auffällig ungleichmäßig verhält. Verschiedene Verfahren dazu messen die sogenannte *Streuung* oder *Dispersion*[11] (vgl. für eine ausführliche Übersicht und Diskussion Gries 2008). Diese wird entweder als zusätzliches Maß angegeben oder fließt in eine korrigierte Häufigkeitsangabe. Normalerweise wird angenommen, dass eine gleichmäßigere Verteiltheit besser ist, also bei den Maßen belohnt wird, eine ungleichmäßigere Verteiltheit wird bestraft.



(a) (b) (c)

Abb. 5
Illustration zu Dispersion (in Anlehnung an Gries / Ellis 2015)

Gries / Ellis verwenden Grafiken ähnlich zu denen in Abb. 5, um die Idealverteilung in (a) einer schiefen Verteilung in (b) gegenüberzustellen. Dabei wird implizit vorausgesetzt, dass die Teile vom Umfang her der Zusammensetzung einem für die Aufgabenstellung ausgewogenen Korpus entsprechen bzw. die Häufigkeiten dahingehend relativiert sind, intuitiv begründet in Eigenschaften der enthaltenen Sprache anhand z.B. Thema, Quellen, Register usw. Eine Ungleichverteilung wie in Abb. 5 (c) würde genauso behandelt werden wie in (b). Die Abgrenzung der fünf Bereiche und auch ihre quantitativen Anteile lassen sich aber kaum pauschal legitimieren. Wenn die Teile des Korpus für verschiedene Aufgabenstellungen unterschiedliche Relevanz haben, z.B. auch für die Unterstrukturierung anhand der nächsten Lernetappe, müsste jeweils die Korpuszusammensetzung oder die Relativierung angepasst werden. Stellen die fünf Teile etwa Zeitabschnitte dar (von A, den ältesten, bis E, den jüngsten) und soll für eine betont synchrone Betrachtung die Gegenwartsnähe eine größere Rolle spielen, so sollten die Frequenzen gemäß (b) und (c) keineswegs gleichermaßen ‚korrigiert' werden: Die hohen älteren, aber sozusagen ‚verblassten' Frequenzen aus (c) sollten durchaus gedämpft werden, die jüngeren aus (b) aber auf keinen Fall. Beispielhaft wurde eine derartige Modellierung der synchronen Betrachtung in Belica et al. (2010) umgesetzt. Entsprechendes gilt auch, wenn die Teilkorpora Themen oder Registern entsprechen, über die das Lehrmaterial

[11] Das sind Maße, die erfassen, wie weit wie viele Werte vom Durchschnitt entfernt beobachtet werden: Eine geringe Streuung heißt, dass fast alle Werte in der Nähe des Durchschnitts liegen, eine hohe Streuung bedeutet, dass entweder einige ein wenig mehr vom Durchschnitt entfernt sind und/oder dass einzelne sehr weit abweichen.

strukturiert werden soll. Eine Möglichkeit neben der Anpassung der Zusammensetzung der Korpora bzw. der Gewichtung der Beiträge zur Kumulation wäre, ein dynamisches Maß zu entwickeln oder die Dispersion so zu präsentieren, dass die Beiträge der einzelnen Teile ersichtlich sind.

Das einfachste Maß für die Dispersion wäre die Anzahl der Bereiche, in denen ein Wort belegt ist. Ein Wort, das in dem obigen Beispiel in allen fünf Bereichen belegt ist (egal wie oft), erhält dann den Wert 5. Ein Wort, das nur in A oder in E nachgewiesen ist, in den anderen Bereichen aber gar nicht, wird mit 1 bewertet – die Information, in welchem Bereich es vorgefunden wurde, geht verloren. Bildet man die Summe jedoch über aufsteigende Zweierpotenzen, wird eindeutig dokumentiert, in welchen Bereichen das Wort belegt ist und in welchen nicht: Der erste Fall bekommt den Wert 1, der zweite den Wert 16 (vgl. Abb. 6[12]).

$$\boldsymbol{binsumdisp(wort)} = \sum_{\substack{\forall i: \boldsymbol{wort} \in \\ \boldsymbol{Vocab(bereich_i)}}} \boldsymbol{2^i}$$

$\boldsymbol{2^0 + 2^1 + 2^2 + 2^3 + 2^4}$

$\boldsymbol{1 + 2 + 4 + 8 + 16}$

$\boldsymbol{2^0 + \times + \times + \times + \times}$ **= 1**

$\boldsymbol{\times + \times + \times + \times + 2^4}$ **= 16**

Abb. 6
Dispersion als binäre Summe

Dieses Maß wurde für die Korpusauswertung des Projekts „Wortschatzwissen – Ein Referenzwortschatz für die Sekundarstufe I“[13] zur Verfügung gestellt und in einer kleinen Studie zu Kinder- und Jugendbüchern eingesetzt (vgl. Perkuhn 2020). Es zeigt seine Stärken bei extremen Schieflagen in der Verteilung sowie im niederfrequenten Bereich. Extreme Ungleichverteilungen können zustande kommen: bei (i) zeitlicher Gliederung durch ganz frische Neulexeme oder Kurzzeitwörter, (ii) regionaler Unterteilung durch Ortsbezeichnungen oder Eigennamen (Personen, Vereine, Unternehmen), deren Popularität eine begrenzte Reichweite hat, oder z.B. (iii) literarischen ‚Gattungen‘ durch Besonderheiten im Stil, Register oder der Thematik. Vor allem im letzten Bereich sind die Spezifika aber nicht so radikal ausgeprägt, dass Wörter in anderen Bereichen gar nicht vorkommen, sodass die binäre Summe dies nicht erfasst. Dafür wurde eine weitere Beschreibung entworfen, die die Häufigkeiten in allen Bereichen berücksichtigt.

Die Unterteilung bei der bereits erwähnten Studie zum Kinder- und Jugendbuchkorpus orientierte sich an den Altersempfehlungen der jeweiligen Texte. Für jede Sparte getrennt und für das Gesamtkorpus wurde die absolute und die relative Häufigkeit ermittelt (Spalte 2-6 in Tab. 2). Die binäre Summe bewertet die Streuung bei allen vier Formen gleich, da sie in allen vier Teilkorpora vorkommen ($1 + 2 + 4 + 8 = 15$). Dass sich die Formen „ich“ und „sie“ sehr ungleichmäßig verteilen, nahezu komplementär, erkennt man erst beim genauen Auswerten der Häufigkeiten im Vergleich in beide Richtungen.

[12] Um den Summanden „1“ dabeizuhaben, sind die Exponenten alle um eins reduziert. Statt von 1 bis 5 laufen sie von 0 bis 4.
[13] https://www.uni-due.de/germanistik/hass/wortschatzwissen (05.11.2021).

Form	> 7	> 12	> 13	> 14	gesamt	bin. Summe	HK-Code
und	**4.496**	**4.082**	**5.888**	**6.196**	**20.662**	15	9999
	0,0193	*0,0260*	*0,0248*	*0,0261*	*0,0239*		
die	**4.038**	**3.105**	**4.420**	**4.931**	**16.494**	15	9999
	0,0173	*0,0198*	*0,0186*	*0,0208*	*0,0191*		
ich	**1.391**	**2.796**	**6.055**	**3.681**	**13.923**	15	8998
	0,0060	*0,0178*	*0,0255*	*0,0155*	*0,0161*		
sie	**4.901**	**2.143**	**2.421**	**2.738**	**12.203**	15	9888
	0,0210	*0,0137*	*0,0102*	*0,0115*	*0,0141*		

Tab. 2
Die vier häufigsten Formen im Kinder- und Jugendbuchkorpus mit Angaben zu Häufigkeiten (fett: absolut, kursiv: relativ) und Verteilung (in Anlehnung an Perkuhn 2020)

In einem kurzen, gut lesbaren Code soll die Information zur Streuung intuitiv zum Ausdruck gebracht werden. Die Verkettung der absoluten Häufigkeiten ist dafür kein guter Kandidat, da die Teilkorpora unterschiedlich groß sind. Relative Häufigkeiten, besser noch der Rang – wenn quasi pro Spalte nach Häufigkeiten absteigend sortiert würde – wären denkbar, erfordern aber aufgrund ihrer Stelligkeiten ein zusätzliches Trennzeichen. In der Korpuslinguistik übliche Häufigkeitsklassen (HK) stellen ebenfalls ein sehr stabiles Maß für den Vergleich dar. Sie geben den logarithmischen Abstand zum häufigsten Element an, üblicherweise über den Zweierlogarithmus sozusagen die Anzahl der Verdopplungsschritte, bis die maximale Häufigkeit überschritten wird. Ist die Basis des Logarithmus fest vorgegeben, so führt dies bei unterschiedlich großen Korpora mit deutlich unterschiedlichen maximalen Häufigkeiten zu unterschiedlich vielen Häufigkeitsklassen. Bei einem Vergleich der größeren Korpora geschriebener Sprache zu denen gesprochener Sprache am IDS (vgl. Meliss et al. 2018) wird das Vokabular des kleineren Korpus z.B. in nur 15 Klassen eingeteilt statt 31 im Größeren. Beim Versuch, die Häufigkeitsklassenangaben der beiden Korpora unmittelbar aufeinander zu beziehen, wirkt dies im hochfrequenten Bereich plausibel (vgl. Abb. 7, oben links), im niederfrequenten Bereich gehen die Angaben allerdings auseinander. Etwas, was wir im intuitiven Sinne gleich einordnen würden, bekommt unterschiedliche HK-Werte, die nur einmal vorkommenden Hapaxe z.B. den jeweils größten Wert. Sie machen ca. 45-50% des Vokabulars aus. Seltene Ereignisse, gerade auch die Hapaxe, sind aus statistischer Sicht ein undankbarer Effekt der Stichprobenziehung: Einige wenige sind unter- oder gerade richtig repräsentiert, die allermeisten sind überrepräsentiert, ihre relative Häufigkeit in der Grundgesamtheit (was auch immer das beim Phänomen ‚Sprache' ist) ist viel geringer als die in der Stichprobe. Betrachtet man die beiden Korpora als unterschiedlich große Stichproben, die unter den gleichen Bedingungen gezogen wurden, könnten sich einzelne Hapaxe des kleinen Korpus durchaus in derselben Häufigkeitsklasse des größeren Korpus wiederfinden (wären dort also keine Hapaxe mehr, sondern häufiger); durchaus plausibler wäre aber, dass die allermeisten Hapaxe Hapaxe bleiben und dann einer anderen Häufigkeitsklasse zugeordnet würden.

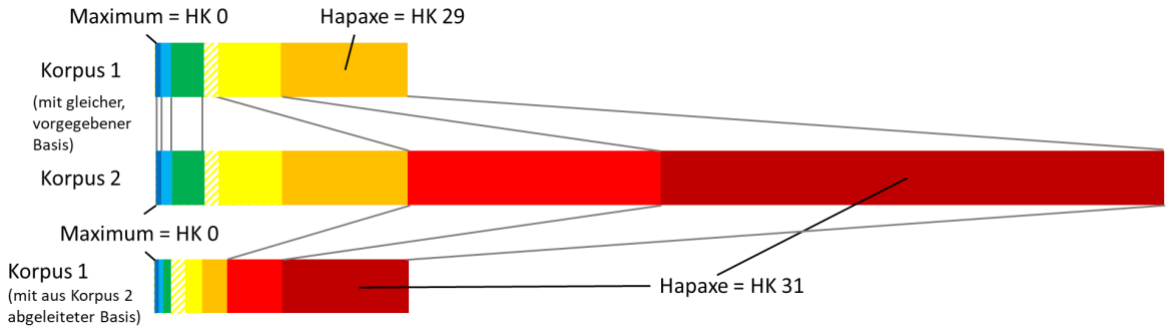


Abb. 7
Logarithmische Häufigkeitsklassen mit fester oder dynamischer Basis

Als Versuch, diesem Umstand gerecht zu werden, ohne die plausible Zuordnung im hochfrequenten Bereich aufzugeben, sollen die Vokabulare beider Korpora auf die gleiche Anzahl logarithmischer Häufigkeitsklassen abgebildet werden. Dazu kann dann aber nicht mehr auf die gleiche Basis zurückgegriffen werden, sondern sie muss dynamisch aus den Eckdaten ermittelt werden. Bei einer Zielvorgabe von *n* Häufigkeitsklassen (wie im Beispiel in Abb. 7: 31) und einer maximalen Häufigkeit von *max* (hier: im Korpus 2) kann man die Basis als *n*-te Wurzel des Maximums berechnen: $b = \sqrt[n]{max}$. Durch die Unteilbarkeit der Hapax-Klasse, die Verschiebung der Klassengrenzen und Rundungsungenauigkeiten muss die Berechnung durch Korrekturterme nachjustiert werden.

Der gleiche Ansatz kann verwendet werden, um von vornherein eine bestimmte Anzahl Häufigkeitsklassen vorzugeben. Für den oben genannten Code soll die Häufigkeitsklasse pro Teilkorpus mit einem einzigen Zeichen repräsentiert werden. Für das kleine Kinder- und Jugendbuchkorpus wurden dafür die Ziffern 1-9 vorgegeben, 0 reserviert für ‚nicht vorhanden', die dann abweichend von der sonstigen Handhabung mit positiver Logik interpretiert werden können: 0 steht für 0 Vorkommen, 1 für Hapaxe, dann aufsteigend bis 9 für Maximum (oder nahe dabei). Für größere Korpora bieten sich die Buchstaben des Alphabets an: Z für 0 Vorkommen, Y für Hapaxe, A für das Maximum.

In der Kodierung der Verteilung in Tab. 2 (letzte Spalte) sieht man auf den ersten Blick, dass das Wort ‚ich' zwar im zweiten und dritten Korpus zur höchstfrequenten Klasse gehört, aber nicht im ersten und vierten. Das Wort ‚sie' ist hingegen nur in der ersten Altersstufe in der höchsten Klasse, sonst erst in der zweithöchsten. Auch wenn beide Wörter in allen Altersstufen in hochfrequenten Bereichen vertreten sind, deutet das auf einen Wechsel der bevorzugten Erzählperspektive hin.

Bei der systematischen Auswertung der HK-Kodierung traten ebenso thematische Schwerpunkte der verschiedenen Altersstufen zutage, überlagert durch geschlechtsbezogene Ausrichtungen. Da bei der Untersuchung Wortformen und nicht Lemmata ausgewertet wurden,

Tab. 3
Unterschiedliche Verwendung von Tempus-Formen

Präsens-Verbformen	**HK-Code**	**Präteritum-Verbformen**	**HK-Code**
fragt	7464	*fragte*	0545
ruft	7353	*rief*	1435
meint	7342	*meinte*	1333
zieht	6453	*zog*	1545
hält	6453	*hielt*	0545
lacht	6352	*lachte*	0434
grinst	6351	*grinste*	0424
erklärt	6343	*erklärte*	0434
denkt	6343	*dachte*	4556
nickt	6325	*nickte*	0545

zeigten sich auch grammatische Phänomene wie der Wechsel der Tempus-Verwendung von der ersten zu den anderen Altersstufen (vgl. Tab. 3).

Alle diese Befunde haben nicht nur geholfen, die Eigenschaften der verschiedenen Teilkorpora zu verstehen. Wenn die Formulierungsentscheidungen von den AutorInnen bewusst oder intuitiv gefällt wurden, um altersgemäß zu formulieren, bieten sie auch Hinweise auf die Elemente anspruchsvollerer und schwierigerer Texte, sowohl inhaltlicher, aber auch sprachlicher Natur, die eventuell auch für die Vermittlung im DaF-Bereich berücksichtigt werden sollten.

Die Streuungskodierung ermöglicht auch, die Häufigkeiten im Gesamtbestand und in den einzelnen Teilkorpora bezüglich der Relevanz für die Aufgabe, z.B. das Vokabular der nächsten Lernetappe, neu zu bewerten. Mit dem in dieser Studie verwendeten Korpus kann eher ein Bezug zum DaM-Unterricht hergestellt werden, da Eigenschaften von Texten herausgearbeitet wurden, die von der Altersempfehlung eher dem Primarbereich vs. dem frühen Sekundarbereich zugeordnet wurden. Inwieweit sich daraus in Analogie eine ähnliche Stufigkeit auf den DaF-Unterricht übertragen ließe, müsste noch sondiert werden. Unabhängig davon ist der Ansatz aber ebenso auf für den DaF-Unterricht zugeschnittene Korpora übertragbar.

4. Kollokationen / Kookkurrenzen[14]

Um eine sprachliche Formulierung einordnen zu können, für welche Lernetappe sie geeignet wäre, ist sie bezüglich erreichtem Lernstand und geplantem Zuwachs zu bewerten (s. Abb. 2). Idealerweise ist ein Großteil der Grammatik und des Vokabulars bekannt und nur in einem von beiden kommt etwas Neues hinzu. Wenn die Formulierungen nicht unter künstlichen Bedingungen mit diesen Vorgaben produziert werden, stellt sich bei natürlich entstandenen und ebenso wirkenden Texten die Frage, welche Freiheiten es gegeben hätte, den Text so oder anders zu formulieren. Sofern nicht Verständlichkeit als Leitprinzip gilt, im Extremfall ausgerichtet an den Empfehlungen für einfache oder leichte Sprache (vgl. Baumert 2016), orientieren sich TextproduzentInnen oft an der Maxime *variatio delectat*: Da wo es Freiheiten gibt, wie z.B. oft bei einer Benennung, werden diese genutzt. Sind mehrere Texte inhaltlich gleichermaßen für die nächste Lernetappe geeignet, können sie vergleichend bewertet werden, wie diese Freiheiten im Verhältnis zum Lernstand, insbesondere aber auch zum geplanten Lernzuwachs stehen. Um bei dem Kriterium Häufigkeit zu bleiben, hieße das, dass Texte, die bei den Freiheiten häufigere, somit ‚wichtigere' Wörter gewählt haben, in die engere Wahl kommen. Ob sich die Häufigkeiten dabei im Bereich der Progression des Lernstandes bewegen, ist allerdings nicht zwingend vorauszusetzen. Hinzu kommt, dass viele vermeintliche Freiheiten bei der Formulierung sich drastisch reduzieren oder sogar ganz auflösen, da für runde und muttersprachlich wohlklingende Formulierungen bestimmte Konstruktionen klar bevorzugt werden.

[14] Die beiden Ausdrücke (und auf andere Art auch ihre englischen Pendants) stehen für verschiedene Ansätze, bestimmte Typen von Wortverbindungen einerseits qualitativ, andererseits quantitativ, zu erfassen. Ungeachtet der Unterschiede in der Perspektive spricht vieles für eine moderne Auslegung, dass sich die beiden Begriffe sehr nahe stehen, evtl. sogar unter idealen Voraussetzungen nahezu konvergieren können. Die Verwendung der beiden Ausdrücke in Kombination zielt darauf, nicht einen der beiden, auf Opposition ausgerichteten Standpunkte einzunehmen und damit nicht die Unterschiede, sondern den gemeinsamen Nenner herauszustellen.

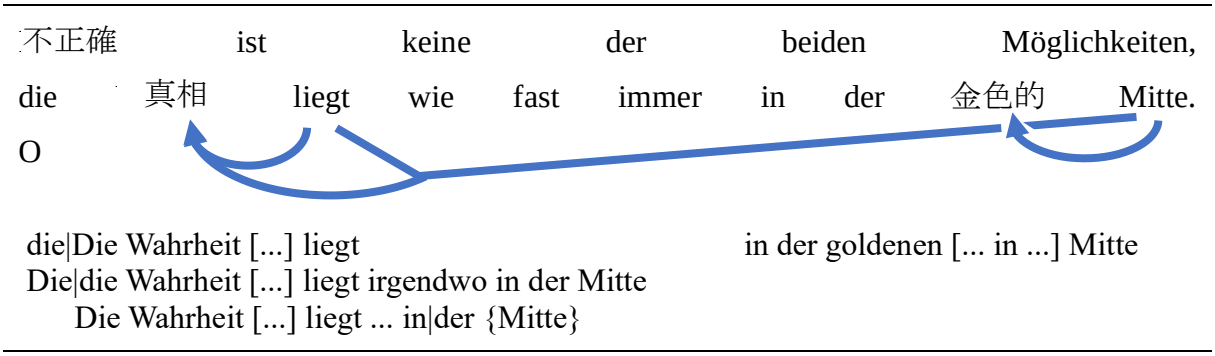


Abb. 8
Übliche Ergänzungen der Syntagmatik als besonders erwartbare Kookkurrenzen

Wenn z.B. der Satz oben in Abb. 8 in die engere Wahl für eine Lernetappe gekommen ist, die mit chinesischen Schriftzeichen überdeckten Wörter[15] aber noch nicht Teil des Lernstandes sind, können wir nicht mit Häufigkeiten für den Lernzuwachs argumentieren, da die Wörter zu selten sind. Wir könnten den ‚Joker' ziehen und uns für den Text und die zu lernenden Wörter entscheiden, da er unabhängig von anderen Kriterien ideal geeignet ist. Eine gute Begründung wäre aber durchaus, dass es sich um die üblich(st)e Formulierung handelt, so üblich, dass sie von MuttersprachlerInnen quasi vorhergesagt werden können. Damit sich DaF-Lernende das Wort aus dem Kontext erschließen können, benötigen sie sicher ein wenig Hilfe.

Die Formulierung kann so verfestigt sein, dass sie als idiomatisch im weiteren Sinne gelten kann, sodass sie *per se* lernenswert ist. Oder sie ist für einen inhaltlichen Aspekt der nächsten Lernetappe die naheliegende Ausdrucksweise und deswegen nicht weniger lernenswert.

Mit dem Nebeneinanderstellen dieser beiden Sichtweisen ist gewissermaßen eine lange Diskussion zu dem Phänomen zusammengefasst. Hausmann (1984) hat in der Germanistik für eine sehr enge Auslegung der Kombinationsaffinität den Begriff ‚Kollokation' eingeführt, die als Übersetzungsauffälligkeit in Erscheinung treten kann, grundsätzlich aber als nicht-freie produktive Kombinatorik eingeschränkt wurde. Mittlerweile hat sich aber die weichere Sicht etabliert (vgl. Hausmann 2004), wie sie in der angelsächsischen Sicht auf Idiomatik schon länger üblich war (vgl. Palmer 1933) und besonders gut von Bahns auf den Punkt gebracht wurde:

> Betrachtet man dieses spezielle Abgrenzungsproblem dagegen aus *fremdsprachendidaktischer* Sicht, verliert es in erheblichem Maß an Bedeutung. Für die fremdsprachliche Wortschatzarbeit ist es vergleichsweise unwichtig ob Wortkombinationen […] Kollokationen oder freie Kombinationen sind – entscheidend ist vielmehr, da[ss] überhaupt die Isolierung des Einzelworts sowie die Einwortgleichung […] überwunden werden und da[ss] Wörter mit (einem oder auch mehreren) Partnern zusammen gelernt werden. (Bahns 1997: 48)

Belica / Perkuhn (2015) sowie Perkuhn (2016) stellen die unterschiedlichen Sichtweisen auf das Phänomen ‚Kollokation' dar und stellen auch den Bezug zur empirischen Ermittlung üblicher Formulierungen her. Mithilfe eines Kookkurrenzanalyseverfahrens lassen sich – bezogen jeweils auf den zugrunde gelegten Datenbestand eines Korpus – übliche Formulierungen ‚berechnen' (vgl. Evert 2008). Als besonders auffällige Kombinationen treten dabei oft besonders idiomatische Wendungen

[15] Falls LeserInnen zufälligerweise die Schriftzeichen kennen sollten, mögen sie sich bitte kurz in die Rolle der Unwissenden versetzen: Ungefähr so ergeht es Lernenden, die die übrigen Wörter gelernt haben, diese Zeichen zwar als sprachliche Zeichen erkennen, sie aber sonst nicht, schon gar nicht in ihrer Bedeutung einordnen und somit auch nicht den Satz in Gänze verstehen können.

(wie z.B. Sprichwörter) in Erscheinung, deren Status in der Sprachvermittlung gesondert betrachtet werden müsste. Sie zeigen aber auch musterhafte übliche Formulierungen wie im unteren Bereich von Abb. 8[16]. Das hierfür verwendete Verfahren (vgl. Belica 1995) zeichnet sich im Vergleich zu anderen dadurch aus, dass es sowohl Freiheiten der Wortstellung und die morphologische Vielfalt der deutschen Sprache berücksichtigt, als auch ein syntagmatisches Muster anbietet, das signifikante Kollokate und häufige Ergänzungen miteinander verbindet.

Eine für didaktische Zwecke geeignete Gesamtliste von typischen Formulierungen bräuchte allerdings noch einen Bezugspunkt, zumindest für Sortier- und/oder Filtermöglichkeiten.

5. Die Wahrheit?

Frequenzen *oder* Kollokationen? Falsch ist keine der beiden Möglichkeiten. Die Wahrheit für die Bewertung der Angemessenheit sprachlicher Formulierungen für einen bestimmten Lernstand liegt wie fast immer in der goldenen Mitte[17]: Es ist genauso wenig sinnvoll, sich jenseits des ‚Jokers' rein an Frequenzen zu orientieren, wie ausschließlich auf die auffälligsten Kookkurrenzen zu achten. Selbst wenn die Auffälligkeiten in einem Korpus ermittelt werden, das ausschließlich aus Texten besteht, die einer Lernetappe entsprechen, fehlt die Binnendifferenzierung, insbesondere die Abgrenzung zu idiomatischen Wendungen im engeren Sinne. Abgesehen davon stehen derartige Korpora mit ausreichendem Umfang leider nicht in absehbarer Zeit zur Verfügung, was die Voraussetzung für die Anwendung des Analyseverfahrens wäre. Für kleinere Lernschritte gilt dies umso mehr. Auch wenn wir auf für entsprechende Zielgruppen adressierte Texte und/oder von ExpertInnen entsprechend eingestufte Texte zurückgreifen könnten, werden die Textmengen naturgemäß klein bleiben – was leider den Bezug zu Frequenz(schicht)en und darauf aufbauend auch die Einschätzung der auffälligen Verbindungen quasi unmöglich macht. Typische ‚flüssige' Verwendungen können durchaus Elemente enthalten, die auch in sehr großen Korpora weniger frequent sind, in sehr kleinen Korpora dann für die Metriken sozusagen unter dem Radar bleiben. Auch die ersatzweise Untersuchung des Kinder- und Jugendbuchkorpus stößt dabei an ihre Grenzen.

Für ein größeres Lernziel, wie z.B. das Erreichen der Stufe B2 des Europäischen Referenzrahmens, können wir uns aber eventuell mit einem Trick behelfen: Mit einem Zeitungskorpus können wir einen Weg skizzieren, wie die Auswahl der lernenswerten Formulierungen für das (Teil-)Ziel „Zeitung lesen können" bestimmt wird. Aus der Menge aller auffälligen Verbindungen zu einer noch festzulegenden Lemmastrecke werden diejenigen Formulierungen herausgefiltert, die entweder ganz – oder zu einem großen Teil – mit einem vorgegebenen Wortschatz abgedeckt sind. Dabei könnte man so vorgehen, dass alle flektierten Formen des Wortschatzes ggf. vereinigt mit der Lemmastrecke als Positivliste dem Analyseverfahren vorgegeben werden, sodass direkt nur erlaubte Syntagmen ermittelt werden. Alternativ können auch alle Syntagmen zu der Lemmastrecke gebildet und anschließend gefiltert werden. Beide Vorgehen können so umgesetzt werden, dass die Syntagmen ausschließlich oder zu einem bestimmten Anteil aus dem vorgegebenen Wortschatz bestehen. Wir haben uns für die zweite Vorgehensweise entschieden, da uns die Syntagmen in einer Kookkurrenzdatenbank (Belica 2007)[18] bereits vorliegen. Zur Lemmastrecke des elektronischen Valenzwörterbuchs E-

[16] Rot hervorgehoben sind die Suchausdrücke, blau die auffälligen Partner, weitere Wörter sind vereinfachte Darstellung des ermittelten syntagmatischen Musters.
[17] … nur als Tipp, um auch das letzte unbekannte Wort aus Abb. 8 aufzulösen.
[18] http://corpora.ids-mannheim.de/ccdb/ (05.11.2021).

Valbu[19] wurden die Syntagmen ermittelt, die größtenteils durch den Wortschatz der Stufe B2 abgedeckt sind, und als ‚Kollokationsschatz' DeReKoll publiziert (vgl. Perkuhn et al. 2015).

Zahl ... liegt deutlich höher
dürfte|die|tatsächliche **Zahl** [noch|aber ...] höher liegen
Dunkelziffer ... deutlich höher liegen
die|Die **Dunkelziffer** liegt weitaus|weit höher
die **Dunkelziffer** liegt [...] weit höher
lägen weit ... den **Grenzwerten**

Legende:
blau: Kollokate,
schwarz: vereinfacht dargestellte Ergänzungen,
fett: Substantiv,
unterstrichen: nicht in vorgegebenem Wortschatz

Abb. 9
Gefilterte und markierte Syntagmen zum Verb *liegen* in DeReKoll[20]

Bei den Beispielen in Abb. 9 sieht man, dass die Syntagmen in den ersten beiden Zeilen vollständig aus dem vorgegebenen Wortschatz stammen; fett hervorgehoben sind Elemente, die eine geforderte Eigenschaft für das Syntagma erfüllen, in diesem Fall das Vorhandensein eines Substantivs. Die weiteren vier Zeilen zeigen sehr ähnliche Syntagmen, allerdings mit (unterstrichenen) Elementen, die nicht im vorgegebenen Wortschatz enthalten sind. Egal, ob dieser Wortschatz sich rein an Häufigkeiten orientiert oder durch einen anderen Kontext (wie die Expertise zu Referenzniveaus) vorgegeben ist, sollten diese Wörter als Kandidaten betrachtet werden, die zum Wortschatz hinzugenommen werden können. Die Entscheidung darüber liegt allerdings im Ermessen, inwieweit sie zur thematisch-diskursiven Ausrichtung des nächsten Lernschrittes bzw. der Lernetappe sinnvoll erscheinen.

Auch wenn die Auswahl und die Darstellungsvarianten sich hier stark an der Vorgabe orientiert haben (und das Korpus durch Aussagen zu Verbrechensstatistiken und zur Umweltbelastung geprägt ist), sollte die Übertragbarkeit auf diverse andere und allgemeinere Szenarien evident sein, genauso allerdings wie die noch zu bewältigenden Herausforderungen und eine systematische Evaluierung vor dem Anwendungshintergrund.

[19] https://grammis.ids-mannheim.de/verbvalenz (05.11.2021).
[20] http://corpora.ids-mannheim.de/DeReKoll/ccdb/v1.0/ (05.11.2021).

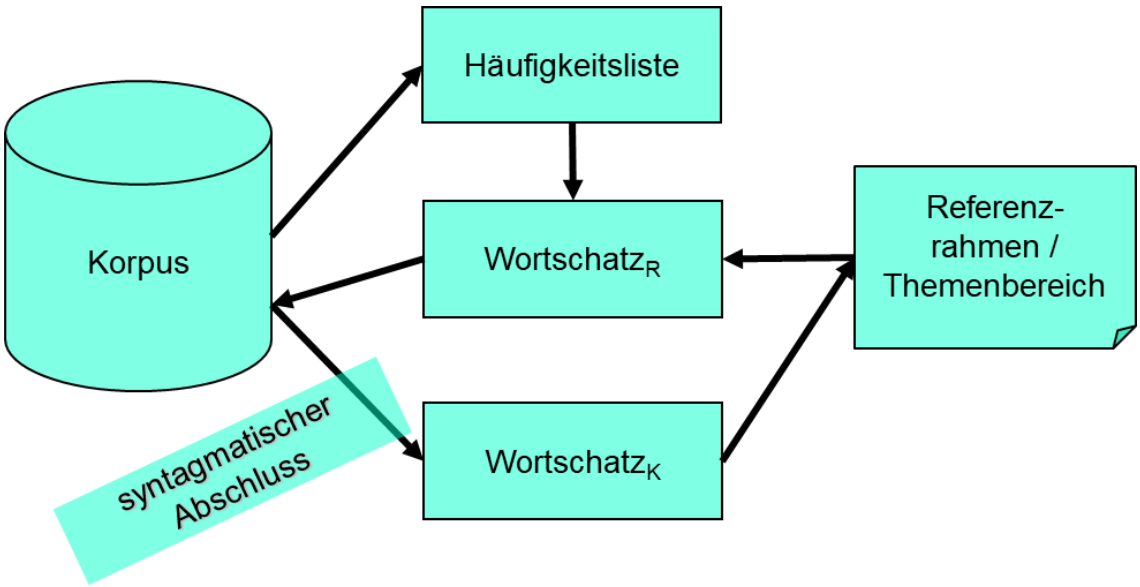


Abb. 10
Einfaches Modell zum Fortschreiben des zu lernenden/lehrenden Wortschatzes

Zusammenfassend lässt sich die Idee wie in Abb. 10 dargestellt skizzieren: Ein initialer Zustand des Referenzwortschatzes $Wortschatz_R$ entsteht durch den Abgleich eines eventuell thematisch strukturierten Referenzrahmens mit einer Wortliste, die entweder über reine Häufigkeitsangabe oder der Berücksichtigung der Dispersion aus einem Korpus ermittelt wurde. Die Dispersion käme vor allem dann zum Tragen, wenn das Korpus in seiner Gesamtheit thematisch nicht unbedingt passend ist, aber eine auswertbare Unterstrukturierung bietet. Über diesen Wortschatz werden Syntagmen ermittelt. Diese werden wiederum als Kollokat(wort)schatz $Wortschatz_K$ mit der Schwerpunktsetzung des Referenzrahmens abgeglichen, insbesondere bei nicht kompletter Abdeckung hinsichtlich des zusätzlich erforderlichen Materials. Der daraufhin erweiterte Referenzwortschatz kann erneut auf den syntagmatischen Abschluss geprüft werden, inwieweit für die vorgeschlagene Syntagmatik zusätzliche Wörter benötigt werden, sodass stufige Erweiterungen entstehen. Ob bzw. wie oft das Verfahren iterieren sollte und sich womöglich ein sinnvoller, stabiler Zustand einstellen kann, müsste in einer Studie mit allen beteiligten Disziplinen eruiert werden.

6. Fazit

Die Diskussion von Wortfrequenzlisten und Kollokationen für den fremdsprachlichen Unterricht hat zumindest im akademischen Bereich eine lange Tradition, nicht zuletzt durch die korpuslinguistische Motivation und Reflexion dieser Konzepte. Auch für die spezielle DaF-Lexikographie sind beide Aspekte dokumentiert (vgl. Zitat in Kap. 2. S. 112/113, sowie Abschnitt 6.6 „Kollokationen“ in Götz 2019, letzterer bis zur Ausgabe von 1993 zurückverfolgbar). Das Lernen und Üben von Vokabeln im bzw. mit Kontext stellt vermutlich unbestritten zumindest eine bereichernde Ergänzung zum traditionellen Lernen dar, für das Französische kann dafür auch auf Lehrmaterial verwiesen werden, dass dieses Vorgehen im Titel trägt (vgl. Fischer / Le Plouhinec 2012). Dass ‚weniger informative‘ Kontexte[21] besser geeignet sind, um Gelerntes für das produktive Beherrschen zu festigen (van den

[21] Leider wird das Kriterium ‚informativ‘ in dem Beitrag nur über subjektive Bewertungen eingeschätzt.

Broek et al. 2018), schränkt die Gültigkeit der Aussage nicht ein. Für den Aspekt des initialen oder früh vertiefenden rezeptiven Erschließens der Bedeutung sind gerade die typischen aussagekräftigen Kontexte relevant, wie es auch in der genannten Studie bestätigt wird. Darüber hinaus ist fraglich, inwieweit Ergebnisse anhand von monosemen Substantiven, für die es ein 1:1-L1-Äquivalent gibt, auf andere Fälle übertragen werden können, insbesondere, wenn Polysemie und weichere Bedeutungen bei Verben und Adjektiven mit feiner Verwendungsdifferenzierung ins Spiel kommen, sowie fest(er)e Fügungen wie Kollokationen i.e.S., bei denen es nicht nur um Form-Bedeutung-, sondern auch um Form-Form-Assoziationen geht.

Schwerpunkt dieses Beitrags war die Darstellung von möglichen Kriterien, anhand derer Lehrmaterial strukturiert werden kann. Worthäufigkeiten zur Strukturierung des Wortschatzes für Lernetappen sind ein erster Ansatz. Für flüssige Formulierungen reicht es aber oft nicht, die nächsthäufigen Wörter hinzuzunehmen. Es sind z.T. sogar eher seltene Wörter, die einen Ausdruck muttersprachlich abrunden. Diese bekommt man nur heraus, wenn man die – bezogen auf den jeweiligen Datenbestand – typischen Formulierungsmuster ermittelt, z.B. mit einer Kookkurrenzanalyse, all das immer in Relation zum Sprachstand bzw. anvisierten Lernfortschritt.

Bisher wurde dieses Gedankengerüst nur exemplarisch angewendet. Für weitergreifende Einsatzszenarien müsste disziplinübergreifend erarbeitet werden, in welchen Daten mit welcher Unterteilung welche sprachlichen Einheiten (alleine oder in Kombination) wie bewertet werden sollten, sodass die Metriken für allgemeingültige oder spezielle Fragestellungen am sinnvollsten kalibriert werden.

Literatur und Ressourcen

Ahrenholz, Bernt / Wallner, Franziska (2013): Digitale Korpora und Deutsch als Fremdsprache. In: Ahrenholz, Bernt / Oomen-Welke, Ingelore (Hrsg.): *Deutsch als Fremdsprache* (Deutschunterricht in Theorie und Praxis, Bd. 10). Baltmannsweiler: Schneider Verlag Hohengehren, 261-272.

Bahns, Jens (1997): *Kollokationen und Wortschatzarbeit im Englischunterricht*. Tübingen: Narr.

Baumert, Andreas (2016): *Leichte Sprache – Einfache Sprache*. Literaturrecherche Interpretation Entwicklung. Open Access. Hannover. https://serwiss.bib.hs-hannover.de/frontdoor/deliver/index/docId/697/file/ES.pdf (1.4.2020).

Belica, Cyril (1995): Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethode. http://corpora.ids-mannheim.de/ (15.12.2020)

Belica, Cyril (2007): *Kookkurrenzdatenbank CCDB - V3*. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. http://corpora.ids-mannheim.de/ (15.12.2020). (alternative direkte URL: http://corpora.ids-mannheim.de/ccdb/)

Belica, Cyril et al. (2007): Web as Corpus: Kooperation mit der Universität Bologna. In: *Sprachreport Sonderheft/März 2007. Auslandskooperationen des Instituts für Deutsche Sprache*. Mannheim, 21-25.

Belica, Cyril et al. (2010): Putting corpora into perspective. Rethinking synchronicity in corpus linguistics. In: Mahlberg, Michaela / González-Díaz, Victorina / Smith, Catherine (Hrsg.): *Proceedings of the Corpus Linguistics Conference 2009*, Liverpool. Liverpool: University of Liverpool.

Belica, Cyril / Perkuhn, Rainer (2015): Feste Wortgruppen/Phraseologie I: Kollokationen und syntagmatische Muster. In: Haß, Ulrike / Storjohann, Petra (Hrsg.): *Handbuch Wort und Wortschatz*. (= Handbücher Sprachwissen 3). Berlin / Boston: de Gruyter, 201-225.

Clarke, David F. / Nation, Ian S. Paul (1980): Guessing the meanings of words from context: strategy and techniques. In: *System* 8: 3, 211-220.

Europarat (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren und beurteilen.* Hrsg. vom Goethe-Institut, der KMK, der EDK und dem BMBWK und dem ÖSD. Berlin: Langenscheidt.

Evert, Stefan (2008): Corpora and collocations. In: Lüdeling, Anke / Kytö, Merja (eds.): *Corpus Linguistics. An International Handbook*, 58. Berlin: Mouton de Gruyter, 1212-1248.

Fandrych, Christian / Tschirner, Erwin (2007): Korpuslinguistik und Deutsch als Fremdsprache. Ein Perspektivenwechsel. In: *Deutsch als Fremdsprache* 44, 195-204.

Fischer, Wolfgang / Le Plouhinec, Anne-Marie (2012): *Mots et contexte* - Neubearbeitung: Thematischer Oberstufenwortschatz Französisch. Stuttgart: Klett Sprachen.

Gabrielatos, Costas (2018): Keyness Analysis: nature, metrics and techniques. In: Taylor, Charlotte / Marchi, Anna (eds.): *Corpus Approaches to Discourse: A critical review*. Oxford: Routledge, 225-258.

Glaboniat, Manuela / Müller, Martin / Rusch, Paul (2002): *Profile Deutsch*. Berlin et al.: Langenscheidt.

Goethe-Institut (o.J.): Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen. Kap. 3, Abschn. 3.3 (Beschreibung der Gemeinsamen Referenzniveaus). https://www.goethe.de/z/50/commeuro/303.htm

Götz, Dieter (Hrsg.) (2019): *Langenscheidt Großwörterbuch Deutsch als Fremdsprache*. Neubearbeitung. München: Langenscheidt.

Granger, Sylviane (2017): Learner corpora in foreign language education. In: Thorne, Steven L. / May, Stephen (eds.): *Language and Technology*. *Encyclopedia of Language and Education*. Springer International Publishing: Cham, 427-440.

Gries, Stefan Th. (2008): Dispersions and adjusted frequencies in corpora. In: *International Journal of Corpus Linguistics* 13: 4, 403-437.

Gries, Stefan Th. / Ellis, Nick C. (2015): Statistical Measures for Usage-Based Linguistics. In: *Language Learning* 65: S1, 228-255.

Hausmann, Franz Josef (1984): Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. In: *Praxis des neusprachlichen Unterrichts* 31: 1, 395-406.

Hausmann, Franz Josef (2004): Was sind eigentlich Kollokationen? In: Steyer, Kathrin (Hrsg.): *Wortverbindungen – mehr oder weniger fest*. Berlin: de Gruyter, 309-334.

Kaeding, Friedrich Wilhelm (1897/98): *Häufigkeitswörterbuch der deutschen Sprache*. Band 1, 2. Steglitz bei Berlin.

Kegel, Gerd (1974): *Sprache und Sprechen des Kindes*. (rororo studium; 56). Reinbek: Rowohlt.

Klann-Delius, Gisela (1999): *Spracherwerb*. (Sammlung Metzler, Bd. 321). Stuttgart: Metzler.

Kyongho, Hwang / Nation, Ian S. Paul (1989): Reducing the vocabulary load and encouraging vocabulary learning through reading newspapers. In: *Reading in a Foreign Language* 6: 1, 323-335.

Leibniz-Institut für Deutsche Sprache (2009): Korpusbasierte Wortgrundformenliste DEREWO, v-40000g-2009-12-31-0.1, mit Benutzerdokumentation, http://www.ids-mannheim.de/kl/derewo/, Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2009. (direkte URL: http://www1.ids-mannheim.de/fileadmin/kl/derewo/derewo-v-40000g-2009-12-31-0.1.zip)

Leibniz-Institut für Deutsche Sprache (2013/2015): Kollokationsschätze zum Deutschen Referenzkorpus DeReKo, Version E-Valbu. http://corpora.ids-mannheim.de/DeReKoll/ccdb/v1.0/

Leibniz-Institut für Deutsche Sprache (o.J.): E-Valbu: Elektronisches Wörterbuch zur Verbvalenz. https://grammis.ids-mannheim.de/verbvalenz.

Lüdeling, Anke / Walter, Maik (2009): Korpuslinguistik für Deutsch als Fremdsprache. In: *Sprachvermittlung und Spracherwerbsforschung*, 1-37.

Lüdeling, Anke / Walter, Maik (2010): Korpuslinguistik. In: Krumm, Hans-Jürgen et al. (Hrsg.): *Handbuch Deutsch als Fremd- und Zweitsprache*. Berlin: Mouton de Gruyter, 315-322.

McEnery, Tony / Xiao, Richard (2010): What corpora can offer in language teaching and learning. In: *Handbook of Research in Second Language Teaching and Learning*. London / New York: Routledge, 364-380.

Meliss, Meike et al. (2018): Creating a List of Headwords for a Lexical Resource of Spoken German. In: Čibej, Jaka et al. (Hrsg.): *Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts*, 17-21 July, Ljubljana. Ljubljana: Znanstvena založba, 1009-1016.

Mukherjee, Joybrato (2002): *Korpuslinguistik und Englischunterricht. Eine Einführung*. Berlin u.a.: Peter Lang.

Nation, Ian S. Paul (2001): *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, Ian S. Paul (2016): *Making and Using Word Lists for Language Learning and Testing*. Amsterdam: Benjamins,

Palmer, Harold E. (1933): *Second Interim Report on English Collocations*, Submitted to the Tenth Annual Conference of English Teachers under the Auspices of the Institute for Research in English Teaching. Tokyo.

Perkuhn, Rainer (2016): Collocation(s) in German minds. In: Sanromán Vilas, Begoña (Hrsg.): *Collocations cross-linguistically. Corpora, dictionaries and language teaching.* Helsinki: Société Néophilologique, 167-192.

Perkuhn, Rainer (2020): Überlegungen zur sprachstandbezogenen Relativierung von Wortschätzen. Ein theoretischer Rahmen und eine kleine empirische Studie. In: Gür-Şeker, Derya (Hrsg.): *Wörter, Wörterbücher, Wortschätze. (Korpus-)Linguistische Perspektiven*. Duisburg: Universitätsverlag Rhein-Ruhr, 194-213.

Perkuhn, Rainer / Belica, Cyril (2006): Korpuslinguistik – das unbekannte Wesen. oder Mythen über Korpora und Korpuslinguistik. In: *Sprachreport* 1/2006. Mannheim, 2-8.

Perkuhn, Rainer et al. (2015): Valenz und Kookkurrenz. In: Domínguez Vázquez, María José / Eichinger, Ludwig M. (Hrsg.): *Valenz im Fokus.* Grammatische und lexikographische Studien. Festschrift für Jacqueline Kubczak. Mannheim: Institut für Deutsche Sprache, 175-196.

Scott, Mike / Tribble, Christopher (2006): *Textual Patterns: Key words and corpus analysis in language education.* Studies in Corpus Linguistics 22. Amsterdam / Philadelphia: John Benjamins.

Sinclair, John (1991): *Corpus, Concordance, Collocation*. London: Oxford University Press.

Sinclair, John (ed.) (2004): *How to use corpora in language teaching*. Studies in Corpus Linguistics 12. Amsterdam / Philadelphia: John Benjamins.

Sparck Jones, Karen (1971): *Automatic keyword classification for information retrieval.* London, England: Butterworths.

Tschirner, Erwin (2005): Korpora, Häufigkeitslisten, Wortschatzerwerb. In: Heine, Antje / Hennig, Mathilde / Tschirner, Erwin (Hrsg.): *Deutsch als Fremdsprache – Konturen und Perspektiven eines Fachs*. München: Iudicium, 133-149.

van den Broek, Gesa S. E. et al. (2018): Contextual Richness and Word Learning: Context Enhances Comprehension but Retrieval Enhances Retention. In: *Language Learning* 68: 2, June 2018, 546-585.

Wallner, Franziska (2013): Korpora im DaF-Unterricht – Potentiale und Perspektiven am Beispiel des DWDS. *Revista Nebrija de Lingüística Aplicada* 13. https://www.nebrija.com/revista-linguistica/korpora-im-daf-unterricht-potentiale-und-perspektiven-am-beispiel-des-dwds.html (01.04.2020).

Kurzbio:
Rainer Perkuhn, seit 2002 am Leibniz-Institut für Deutsche Sprache, Mannheim, im Programmbereich Korpuslinguistik. Davor tätig in Forschung / Lehre an den Universitäten Bielefeld, Duisburg, Karlsruhe; Lehraufträge an den Universitäten Freiburg, Mannheim, Göttingen, Heidelberg. Diverse Publikationen und Workshops zu korpuslinguistischer Methodik und Anwendungsbezügen, Schwerpunkte: Operationalisierung von Wortbegriffen u.a. für Frequenzbestimmungen, sowie Kookkurrenzanalyse.

Anschrift:
Rainer Perkuhn
Programmbereich Korpuslinguistik
Leibniz-Institut für Deutsche Sprache
Mannheim
perkuhn@ids-mannheim.de