

Analysemethode und Datengrundlage können die Ergebnisse beeinflussen: Selektiver Einfluss der Stammfrequenz für *ver*-präfigierte Verben

Julian Heister[‡], *Luise Bartels*, *Judith Heide*[‡], *Kay-Michael Würzner*[‡]
Universität Potsdam

1. Stamm- und Vollformfrequenzeffekte für *ver*-präfigierte Verben: Die Ausgangsstudie von Bartels et al. (2008)

Bartels et al. (2008; vgl. auch Bartels 2009 und Heide et al. 2008) untersuchten den Einfluss von Stamm- und Ganzwortfrequenz auf die Verarbeitung von deutschen *ver*-präfigierten Verben. Der Wortstamm der komplexen Wörter war entweder ein Nomen (*ver-pflich-ten*), ein Verb (*ver-meiden*) oder aber ambig in Bezug auf den Nomen/Verb-Status (*ver-handeln*). Es wurde ein Experiment zum visuellen lexikalischen Entscheiden durchgeführt, wobei die *ver*-Verben morphologisch geprimit wurden. Als Prime wurden relatierte und unrelatierte Wortstämme (z.B. HANDEL bzw. SUCHE für *verhandeln*) oder Kombinationen aus Wortstamm und Infinitivendung *-en* verwendet (z.B. HANDELN bzw. SUCHEN). Die Primewörter wurden nur sehr kurz präsentiert (62 ms) und waren maskiert, so dass eine bewusste Verarbeitung nicht möglich war. Zur statistischen Auswertung der erhobenen Reaktionszeitdaten führten Bartels et al. (2008) zwei Varianzanalysen (ANOVA) durch. In einer ANOVA wurden 130 Wörter mit ähnlicher Stammfrequenz, aber hohen vs. niedrigen Ganzwortfrequenzen analysiert und in einer zweiten ANOVA erfolgte die Analyse von 170 Wörtern mit hohen vs. niedrigen Stammfrequenzen bei ähnlicher Ganzwortfrequenz.

Die Arbeit dieser Autoren wurde von der Deutschen Forschungsgemeinschaft unterstützt ([‡]DFG KL 955/12-1 bzw. [‡]DFG BU 1420/3-1).

Die Autoren berichten folgende Ergebnisse: 1. Maskierte, *relatierte* Primes beschleunigen die visuelle Worterkennung. 2. Sowohl die Ganzwort- als auch die Stammfrequenz beeinflussen die Reaktionszeiten. In der ANOVA lassen sich diese Effekte allerdings nur in der Subjekt- nicht aber in der Itemanalyse nachweisen. 3. Die Wortart des Stamms beeinflusst die Erkennung von *ver-*präfigierten Verben. Wörter mit Verbstamm wurden am schnellsten erkannt und Wörter mit Nomenstamm am langsamsten. 4. Es besteht eine Interaktion von Ganzwortfrequenz und Priming. Niedrigfrequente Wörter profitieren stärker vom morphologischen Priming als hochfrequente Wörter. Laut Bartels et al. (2008) sind die Ergebnisse am besten im Rahmen von sog. Zwei-Routen-Modellen (Caramazza et al. 1988, Frauenfelder & Schreuder 1992; Baayen & Schreuder 1999) erklärbar. Diese Modelle nehmen an, dass morphologisch komplexe Wörter sowohl ganzheitlich als auch morphembasiert verarbeitet werden können und kombinieren damit die gegensätzlichen Ansätze der obligatorischen morphologischen Dekomposition (Taft & Forster 1975, Taft 2004) und der ganzheitlichen Auflistung (Butterworth 1983). Beispielsweise sagen Zwei-Routen-Modelle voraus, dass hochfrequente Wörter ganzheitlich im mentalen Lexikon gespeichert sind, während niedrigfrequente Vollformen morphembasiert verarbeitet werden. Diese Annahme kann erklären, warum niedrigfrequente Vollformen stärker von morphologischem Priming profitieren als hochfrequente Wörter. Da niedrigfrequente Wörter im mentalen Lexikon als Morpheme vorliegen, ist der Prime identisch mit dem Stamm-Morphem. Hochfrequente Wörter sind dagegen ganzheitlich gespeichert und morphologisches Priming daher weniger wirksam.

Während die Ergebnisse von Bartels et al. (2008) schlüssig interpretiert wurden, lässt sich das methodische Vorgehen aus mindestens drei Gründen verbessern: Erstens korrelieren die Ganzwort- und Stammfrequenzen der 182 verwendeten Wörter miteinander (CELEX: $r=.32$). Eine getrennte Untersuchung der beiden Parameter erscheint daher problematisch. Zwar gleichen Bartels et al. Stamm- bzw. Ganzwortfrequenzen an, dennoch bleibt die Korrelation auch innerhalb des Experimentalmaterials bestehen ($r=.27$ bzw. $r=.38$). Zweitens liefert die ANOVA in Subjekt- und Itemanalyse

unterschiedliche Ergebnisse, so dass eine stringente Interpretation erschwert wird. Darüberhinaus verliert die ANOVA durch den notwendigen Ausschluss von Daten an statistischer Power. Drittens sind die aus dem CELEX-Korpus (Baayen et al. 1990) verwendeten Werte keine tatsächlichen Stammfrequenzen sondern Lemmafrequenzen.

Daher reanalysieren wir die Daten aus Bartels et al. (2008) 1. durch Einbeziehung numerischer Frequenzwerte (anstelle der Einteilung in zwei Frequenzgruppen), 2. durch Verwendung der dlex-Datenbank (Heister et al. eingereicht) und 3. durch die Analyse mit linear gemischten Modellen (LMM; anstelle einer Varianzanalyse). Die Verwendung von LMM erforderte 4. eine Transformation der Reaktionszeiten (vgl. Abschnitt 2.4). Inhaltlich führen wir alle Analysen getrennt nach Wortarten des Stamms durch, da bereits Bartels et al. Unterschiede zwischen den Wortarten finden.

2. Reanalyse der Daten

2.1 Numerische statt kategorielle Frequenzen

Bartels et al. (2008) verwenden Frequenz als kategorielle Variable mit den Stufen ‚hoch‘ und ‚niedrig‘, da dies eine der Voraussetzungen für die Durchführung einer Varianzanalyse ist. LMMs versetzen uns in die Lage, Frequenzen in ihrer eigentlichen Form als numerische Variablen zu verwenden (siehe Abschnitt 2.3). Hiermit berücksichtigen wir die im vorliegenden Datensatz bestehende übliche Korrelation zwischen Stamm- und Ganzwortfrequenz (CELEX: $r = .32$; dlex: $r = .48$). Zusätzlich ermöglichen die numerischen Frequenzen einen Vergleich zweier lexikalischen Datenbanken (CELEX und dlex).

Neben der Verwendung von numerischen Frequenzen wurden alle Frequenzwerte um 1 erhöht und dann log-transformiert, um den Einfluss von Ausreißern vor allem im höherfrequenten Bereich zu verringern (Keene 1995).

2.2 dlex statt CELEX

dlex (<http://dlexdb.de>) ist eine sich in der Entwicklung befindende lexikalische Datenbank, deren Grundlage das Digitale Wörterbuch der Deutschen Sprache (DWDS, Geyken 2007) bildet. Das DWDS enthält 2.200.000 Types (120 Millionen Token) und ist damit sehr viel umfangreicher als CELEX, dessen deutscher Teil 360.000 Types (5.4 Millionen Token) enthält. Während CELEX für 88 der 182 Wörter aus Bartels et al. (2008) als Ganzwortfrequenz den Wert null angibt, liefert dlex nur für 17 Wörter kein Ergebnis. Für die Stammfrequenz hat CELEX für 5 Wörter keinen Eintrag, während dlex jedem Wort eine Stammfrequenz zuordnen kann.

Grundsätzlich ermöglicht dlex die Zusammenstellung von gut kontrolliertem Sprachmaterial (z.B. für Experimente oder Therapiestudien). Gleichzeitig können für bestehende Wortlisten verschiedene Variablen abgefragt werden. Neben der Ganzwort- und Stammfrequenz lassen sich über dlex eine ganze Reihe weiterer Normwerte abrufen, z.B. Wortanfangsfrequenzen, Dokumentfrequenzen und Wortnachbarschaftsmaße (siehe auch Heister et al., eingereicht).

Neben seiner Größe hat dlex gegenüber CELEX den Vorteil, dass Lemma- und Stammfrequenzen unterschieden werden. Abbildung 1 veranschaulicht die unterschiedlichen Frequenzberechnungen von dlex und CELEX am Beispiel von *verglasen*. Die Stammfrequenz für GLAS berechnet sich in dlex aus der Addition der Frequenzen von *Glas*, der Pluralflexion *Gläser*, den verschiedenen Kasusflexionen (z.B. *Glases*, *Gläsern*) sowie allen von GLAS abgeleiteten Derivationsformen (z.B. *verglasen*, *Glaser*, *glasig*). CELEX hingegen bezieht nur die Frequenz der Grundform und der Flexionsformen mit ein (vernachlässigt also u.a. sämtliche Flexionsformen des derivierten Verbs *verglasen*). Außerdem berücksichtigt dlex bei ambigen Wortstämmen sowohl die Frequenz des zugrundeliegenden Nomens als auch des Verbs, während die von Bartels et al. verwendeten CELEX-Frequenzen nur die Lemmafrequenz des Nomens einbezogen.

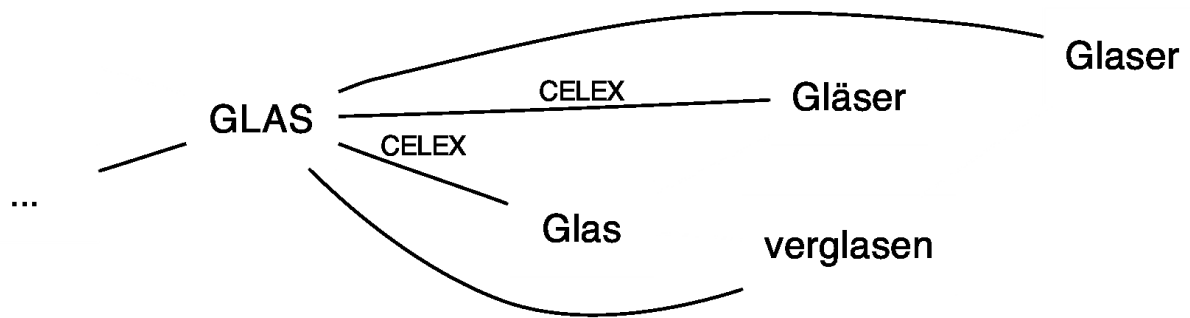


Abbildung 1: Beispiele für Derivationen von „Glas“ in dlex. Beschriftungen für Derivationen in CELEX.

2.3 Linear gemischte Modelle statt Varianzanalysen

Um die unabhängigen Einflüsse von Ganzwort- und Stammfrequenz mit Hilfe von Varianzanalysen zu untersuchen, waren Bartels et al. (2008) gezwungen, zwei *getrennte* ANOVAs für ihren Datensatz aus 182 Wörtern durchzuführen. Einmal wurden die Ganzwortfrequenzen angeglichen und einmal die Stammfrequenzen, um den Einfluss der jeweils anderen Frequenz zu testen. Anstelle von zwei getrennten Analysen lässt sich mit numerischen Frequenzwerten ein linear gemischtes Modell (LMM) formulieren, das gleichzeitig den Einfluss von Ganzwort- und Stammfrequenz schätzt. Außerdem macht ein LMM die Trennung von Subjekt- und Itemanalyse überflüssig, da simultan die Varianzen sowohl zwischen verschiedenen Items als auch zwischen verschiedenen Probanden berücksichtigt werden. In unserem Fall wird die mittlere Reaktionszeit für jede Person und jedes Wort berücksichtigt. Für eine Einführung und weitere Vorteile von LMM gegenüber der linearen Regression siehe Pinheiro & Bates (2000). Da bereits Bartels et al. (2008) Unterschiede zwischen *ver*-Verben mit verschiedenen Wortstämmen berichteten, formulieren wir für jede Wortart des Stamms (Nomen, Verben, ambig) ein eigenes linear gemischtes Model. Untersucht werden der Einfluss von (numerischer) Stamm- und Ganzwortfrequenz, morphologischem Priming sowie die Interaktion von Priming mit (numerischer) Stamm- und Ganzwortfrequenz.

2.4 Reziproke Reaktionszeiten

Betrachtet man die von Bartels et al. (2008) gemessenen Reaktionszeiten, stellt man fest, dass die Unterschiede zu den vom Modell vorhergesagten Reaktionszeiten für langsame Reaktionszeiten größer sind als für schnelle Reaktionszeiten. Diese Beobachtung wird als Heteroskedastizität bezeichnet und verletzt eine der Voraussetzungen für die Anwendung von LMMs. Daher nehmen wir eine Transformation der Reaktionszeiten (RT) vor, indem wir reziproke Werte bilden ($-1/RT$; vgl. Kliegl et al. 2009). Dadurch erreichen wir eine Verringerung der beobachteten Unterschiede.

Die folgenden Ergebnisse beruhen auf der Auswertung von 10466 Reaktionszeiten für 182 unterschiedliche ver-präfigierte Verben und 62 Probanden. Reaktionszeiten über 2000 ms wurden als Ausreißer klassifiziert und von den Analysen ausgeschlossen.

3. Ergebnisse

Tabelle 1 gibt einen Überblick über die untersuchten Effekte der Ganzwort- und Stammfrequenz, des Primings und über die Interaktionen von Frequenz und Priming. Für jede Variable ist ein Koeffizient angegeben, der den Einfluss dieser Variable auf die gemessenen Reaktionszeiten beschreibt. Je höher der Betrag dieses Wertes ist, desto stärker ist der Einfluss der Variable. Der Standardfehler (SF) gibt die Reliabilität des Effektes an. Ist der Betrag des sich ergebenden t-Werts größer als 2.0, ist der Effekt statistisch signifikant.

Tabelle 1:

Effekte der Ganzwort- und Stammfrequenz und des Primings sowie der Interaktionen von Frequenz und Priming für *ver*-präfigierte Verben mit Nomen-, Verb- und ambigem Wortstamm. (Koeffizienten für sechs LMMs, Vergleich von CELEX und dlex, reziproke Reaktionszeiten)

	CELEX			dlex		
	Koef	SF	t	Koef	SF	t
<i>Nomenstamm (n=2863)</i>						
(Intercept)	-1.523	0.041	-36.8	-1.525	0.041	-37.3
GW_F	-0.129	0.064	-2.0	-0.202	0.140	-1.4
ST_F	-0.093	0.039	-2.4	-0.136	0.044	-3.1
Prime	-0.164	0.014	-12.0	-0.164	0.014	-12.0
GW_F x Prime	0.082	0.035	2.3	0.125	0.080	1.6
ST_F x Prime	0.021	0.022	1.0	0.026	0.026	1.0
<i>Verbstamm (n = 4118)</i>						
(Intercept)	-1.622	0.041	-39.9	-1.622	0.041	-39.9
GW_F	-0.096	0.038	-2.5	-0.197	0.072	-2.7
ST_F	-0.018	0.028	-0.7	-0.007	0.029	-0.3
Prime	-0.126	0.011	-11.1	-0.126	0.011	-11.1
GW_F x Prime	-0.009	0.025	-0.4	0.059	0.048	1.2
ST_F x Prime	0.012	0.018	0.7	0.015	0.020	0.7
<i>Ambiger Stamm (n = 3485)</i>						
(Intercept)	-1.639	0.043	-38.3	-1.639	0.043	-38.4
GW_F	-0.122	0.051	-2.4	-0.205	0.115	-1.8
ST_F	-0.043	0.035	-1.2	-0.056	0.043	-1.3
Prime	-0.118	0.013	-9.4	-0.118	0.013	-9.4
GW_F x Prime	0.032	0.030	1.1	-0.072	0.067	-1.1
ST_F x Prime	0.016	0.021	0.8	0.024	0.025	1.0

Erläuterung: GW_F: Ganzwortfrequenz, ST_F: Stammfrequenz, Prime: Einfluss des Primings „x“ bezeichnet Interaktionen, t: t-Wert, t-Werte über 2.0 sind signifikant und fett gedruckt. Intercept: Mittlere reziproke Reaktionszeit für die mittlere Frequenz

Fünf wesentliche Ergebnisse können festgehalten werden:

1. Alle Primingeffekte werden signifikant.
2. Reaktionszeiten für *ver*-Verben mit Verbstamm und ambigem Stamm sind signifikant kürzer als für *ver*-Verben mit nominalem Stamm.
3. Nur für *ver*-Verben mit Nomenstamm ergeben sich Stammfrequenzeffekte.
4. Mit CELEX-Frequenzen wird der Einfluss der Ganzwortfrequenz für alle Wortstämme signifikant. Mit dlex-Frequenzen dagegen nur für Verbstämme.
5. Nur mit CELEX-Frequenzen ergibt sich für *ver*-Verben mit Nomenstamm eine Interaktion von Priming und Ganzwortfrequenz. Alle anderen Interaktionen (für CELEX und dlex-Frequenzen) werden nicht signifikant.

4. Zusammenfassung und Diskussion

Die Ergebnisse der linear gemischten Modelle stützen weitestgehend die Schlussfolgerungen aus Bartels et al. (2008). Die Verwendung intervallskalierter Frequenzen ermöglichte uns, beide ANOVAs aus Bartels et al. simultan in einem LMM zu berechnen. Ein Vorteil der vorgestellten dlex-Datenbank ist die Bereitstellung tatsächlicher Stammfrequenzen im Gegensatz zu den bislang aus CELEX verwendeten Lemmafrequenzen. Im Unterschied zu Bartels et al. wurden die Reaktionszeiten transformiert.

Die Transformation der Reaktionszeiten in reziproke Reaktionszeiten ändert nichts am starken Einfluss des Primings. Wir interpretieren den Primingeffekt wie Bartels et al. (2008) als Nachweis für eine morphologische Zerlegung. Der unabhängige Einfluss der Ganzwortfrequenz spricht für eine ganzheitliche Verarbeitung und bestätigt Bartels et al. (2008), dass morphologisch komplexe Wörter sowohl ganzheitlich als auch morphembasiert verarbeitet werden können. Im Gegensatz zu Bartels et al. ist keine der Interaktionen von Frequenz und Priming über die Frequenznormen von CELEX und dlex stabil. Die Interaktion aus Ganzwortfrequenz und Priming ist nur für *ver*-Verben mit Nomenstamm in CELEX signifikant. Der fehlende stabile Einfluss von Frequenz

auf das Priming in den Analysen mit reziproken Reaktionszeiten lässt keine Aussage darüber zu, ob hochfrequente Wörter ganzheitlich und niedrigfrequente Vollformen morphembasiert im mentalen Lexikon gespeichert sind. Allerdings zeigt Bartels (2009) in der Analyse der Ablenker, dass der Primingeffekt kein einfacher orthographischer Ähnlichkeitseffekt ist.

Bezüglich der Wortarten bestätigen unsere Reanalysen, dass *ver*-präfigierte Verben mit Nomenstamm schwieriger zu verarbeiten sind. Bartels et al. finden für diese *ver*-präfigierten Verben neben verlängerten Reaktionszeiten auch eine geringe Antwortgenauigkeit (siehe Bartels 2009). Die Verarbeitung von *ver*-Verben mit Nomenstamm ist im Gegensatz zu anderen *ver*-Verben auch (CELEX) wenn nicht sogar ausschließlich (dlex) von der Stammfrequenz abhängig. In dlex lässt sich dieser unterschiedliche Verarbeitungsmechanismus für Verben mit Nomenstamm deutlicher abbilden als in CELEX. Die besondere Rolle der *ver*-präfigierten Verben mit Nomenstamm sollte in weiteren Untersuchungen näher beleuchtet werden. Der fehlende Einfluss von Stamm- und Ganzwortfrequenz für Verben mit ambigem Stamm in dlex könnte darauf hinweisen, dass Wörter dieser Gruppe entweder wie Verben mit Verbstamm oder wie Verben mit Nomenstamm verarbeitet werden.

Methodisch haben wir verdeutlicht, dass die Auswertung mit linear gemischten Modellen (Bates 2008) und der Vergleich verschiedener Korpora (CELEX und dlex) Ergebnisse psycholinguistischer Experimente erhärten oder präzisieren können. Die Ergebnisse legen außerdem nahe, dass Transformationen der abhängigen Variablen häufiger beachtet werden sollten. Die dlex-Datenbank bietet wegen ihrer Datenvielfalt und echten Stammfrequenzen in Zukunft eine gute Alternative zu CELEX (Heister et al., eingereicht).

5. Literatur

Baayen, H. R., Piepenbrock, R. & Gulikers, L. (1990). *The CELEX Lexical Database (Release 2)* [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995.

- Bartels, L., Heide, J., Lorenz, A. & Burchert, F. (2008). Der Einfluss der Stamm- und Ganzwortfrequenz auf die Verarbeitung von derivierten Verben im Deutschen. In: Wahl, M., Heide, J. & Hanne, S. (Hrsg.) *Spektrum Patholinguistik 1 – Der Erwerb von Lexikon und Semantik*. Universitätsverlag Potsdam. 175-182.
- Bartels, L. (2009). *Der Einfluss von Stamm- und Ganzwortfrequenz auf die Verarbeitung derivierter Verben im Deutschen*. Diplomarbeit. Universität Potsdam.
- Butterworth, B. (1983). Lexical Representation. In: Butterworth, B. (Hrsg.). *Language Production (2): Development, Writing and Other Language Processes*. London: Academic Press. 257–294.
- Caramazza, A., Laudanna, A. & Romani, C. (1988). Lexical Access and Inflectional Morphology. *Cognition* 28 (3), 297–332.
- Geyken, A. (2007). The DWDS Corpus: A Reference Corpus for the German Language of the 20th Century. In: Fellbaum, C. (Hrsg.). *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects*. London: Continuum Press.
- Heide, J., Bartels, L., Lorenz, A. & Burchert, F. (2008). *Lexical decision on German prefixed verbs: Full form and root frequency effects*. Vortrag auf der 6th International Conference on the Mental Lexicon. Banff/ Kanada.
- Heister, J. Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A. & Kliegl, R. (eingereicht). dlex – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*.
- Keene, O. N. (1995). The log transformation is special. *Statistics in Medicine* (14), 811-819.
- Kliegl, R., Masson, M. E. J., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition*, im Druck.
- Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.
- Taft, M. & Forster, K. I. (1975). Lexical Storage and Retrieval of Prefixed Words. *Journal of Verbal Learning and Verbal Behavior* 14 (6), 638–647.

Kontakt

Julian Heister

heister@uni-potsdam.de