Anzeiger für Slavische Philologie
(XXXIV) 2006, 41–74

Graphemhäufigkeiten des Slowenischen (und anderer slawischer Sprachen)

Ein Beitrag zur theoretischen Begründung der sog. Schriftlinguistik

Peter Grzybek / Emmerich Kelih / Ernst Stadlober (Graz)

1. Zusammenfassung

Im vorliegenden Text geht es um die Untersuchung von Graphemhäufigkeiten in slawischen Sprachen. Ausgehend von allgemeinen theoretischen Erörterungen, welche die Untersuchung in einen synergetischen Kontext stellen, wird wesentlich das Spektrum der sog. Schriftlinguistik erweitert. Es wird ein in der jüngsten Vergangenheit an verschiedenen Sprachen mehrfach diskutiertes Verteilungsmodell geprüft und im Hinblick auf mögliche sprachvergleichende bzw. sprachübergreifende Implikationen reflektiert: auszumachende methodologische Probleme der Datenqualität führen zur Notwendigkeit einer grundlegenden Neu-Analyse slowenischer Daten, die erstmals in einen solchen Zusammenhang gestellt werden. Die Ergebnisse weisen auf sprachübergreifende Prinzipien einerseits, sprachspezifische Bedingungen andererseits hin.

2. Buchstabenhäufigkeiten und die sog. „Schriftlinguistik“

Nach wie vor stößt man auch und gerade in der modernen Linguistik auf die Auffassung, die Beschäftigung mit der Vorkommenshäufigkeit von Buchstaben oder Graphemen als Einheiten „niederer Sprachebenen“ sei ein naives Unterfangen: Bestand und Beschaffenheit alphabetischer Systeme – so die Annahme – seien zu sehr durch Zufälligkeiten oder Willkürlichkeiten in Geschichte, Kultur oder Politik geprägt, als dass man hier zu systematischen Einsichten gelangen könne. Diese Befürchtung mag nicht unberechtigt sein, wenn man bedenkt, dass auch heute noch – in der Regel freilich „fachfremde“, insbesondere informatik-basierte – Zugänge zu Sprache durch einen mitunter recht naiven Umgang mit Buchstaben und Graphemen als vermeintlichen Konstituenten von größeren sprachlichen Einheiten (wie etwa von

Wörtern) umgehen. Andererseits aber erweist sie sich selbst als überaus naiv, da ihr eine Reihe von impliziten Voraussetzungen zugrunde liegt, die eine mangelnde Reflexion der Verhältnisse signalisieren. Denn überwiegend werden die Bedenken und Einwände mit einem Verweis auf die in den einzelnen Sprachen sehr unterschiedlichen Graphem-Phonem- bzw. Phonem-Graphem-Relationen begründet; dem liegt die Frage zugrunde, welche Phoneme durch welche Grapheme abgebildet werden, bzw. welche Grapheme zur Abbildung welcher Phoneme dienen. Probleme werden insbesondere in den Fällen (bzw. Sprachen) gesehen, die in der kognitiven Psychologie der letzten Jahre mit dem Stichwort ‚deep orthography' (vs. ‚shallow orthography') bezeichnet worden sind, in denen wir es mit einem Zustand zu tun haben, der weit von einer 1:1-Korrespondenz entfernt ist. Damit wird im Grunde genommen aber eine genuin andere Fragestellung ins Spiel gebracht, nämlich die nach dem Zusammenhang von Schriftlichkeit und Mündlichkeit, wobei sich der Zusammenhang aus linguistischer Sicht in der Regel eher als Primat der Rede, und weniger, wie im poststrukturalistischen Diskurs, als Primat der Schrift darstellt.[1]

Das Schriftsystem wird folglich ausschließlich relational, d.h. in Bezug zum System der Rede gesehen, und nicht als eigenständiges, autonomes System mit eigenen Systembedürfnissen und eigenen Prozessen der Selbst-Regulation. Ebenfalls wird in diesem argumentativen Kontext nur zu gern übersehen, dass auch die Ebene der Phonologie keineswegs ein „positives Faktum" ist, sondern theoriegeleitet und theoriebasiert, also abhängig von der einen (oder anderen) sprachwissenschaftlichen Theorie ist.

In gewissem Sinne ist das anders in der sog. Schriftlinguistik (vgl. Coulmas 1981, Glück 1987, Dürscheid ^{3}2006). Diese hat sich in den letzten Jahren als ein eigener linguistischer Teilbereich etabliert, der Schrift als eine „eigenwertige, voll funktionale Realisierungsform von Sprache" versteht. Doch auch wenn sich das linguistische Interesse hier auf Struktur, Typologie, Geschichte, Erwerb und Verarbeitung von Schriftsystemen richtet, bleibt dabei die theoretische Beschäftigung mit der Vorkommenshäufigkeit graphematischer Einheiten als eines ihrer zentralen Charakteristika vollkommen ausgeblendet. Grund dafür ist ganz offensichtlich, dass nach wie vor der Begriff der Vorkommenshäufigkeit und noch mehr derjenige der Vorkommenswahrscheinlichkeit aus theoriebildenden Bereichen der Linguistik ausgeblendet bleibt – offenbar, weil man sich aufgrund fehlender Konzepte keinen seriösen Einblick in das Funktionieren von Sprache erwartet. Dabei missachtet man

[1] Vgl. etwa die entsprechenden Überlegungen in der *Grammatologie* von Derrida (1974: 29), für den es „kein sprachliches Zeichen, das der Schrift vorherginge", gibt. – Vgl. hierzu jüngst auch Harris (2005).

freilich nicht nur eine zentrale Eigenschaft linguistischer Komponenten – nämlich ihre Frequenz und deren systematisches Verhalten –, sondern vor allem auch die Relevanz dieser Eigenschaft innerhalb des strukturellen Zusammenspiels linguistischer Elemente verschiedener Ebenen, wie sie im Rahmen der synergetischen Linguistik (vgl. Köhler 2005) ausführlich beleuchtet werden.

3. Grapheme: Häufigkeiten und Häufigkeitsmodelle

Ohne also einerseits den elementaren Zusammenhang zwischen Graphematik und Phonologie in Abrede stellen zu wollen, andererseits aber das Wesen alphabetischer Systeme nicht auf diesen funktionalen Zusammenhang reduzieren zu wollen, sondern – durchaus im Sinne einer umfassenderen Schriftlinguistik – als eigenständige linguistische Realisationsform zu verstehen, soll es im vorliegenden Text in Erweiterung bestehender Ansätze um eine spezifische Eigenschaft alphabetischer Systeme, nämlich die Vorkommenshäufigkeit der Elemente – sprich: der Buchstaben bzw. Grapheme eines gegebenen Alphabets – gehen. Da es dabei um allgemeine systemrelevante Eigenschaften von Buchstabeninventaren geht, rückt die spezifische Vorkommenshäufigkeit der einzelnen, individuellen Buchstaben in den Hintergrund; in den Fokus des Interesses rückt statt dessen die Frage, inwiefern die Häufigkeiten einer Buchstabenhäufigkeitsverteilung unabhängig davon, um welche(n) Buchstaben es sich jeweils konkret handelt, allgemeinen Regularitäten der Häufigkeitsorganisation unterliegen. Aus diesem Grunde transformiert man die (absteigende) Häufigkeitsverteilung in eine Rangverteilung, bei der die Frequenz des häufigsten Buchstabens (Rang 1) an erster Stelle, diejenige des am seltensten vorkommenden an letzter Stelle steht. Die entscheidende Frage richtet sich dann darauf, ob die Häufigkeiten der einzelnen Ränge in einer bestimmten Relation bzw. Proportion zu einander stehen, und ob bzw. wie diese Relationen am besten beschrieben werden können.

Dieses Vorgehen ist in jüngerer Vergangenheit wiederholt beschrieben worden (Grzybek, Kelih, 2003a,b; Grzybek, Kelih, Altmann 2004), so dass hier auf eine detaillierte Beschreibung verzichtet werden kann. Zudem ist es im Rahmen einer Reihe einzelsprachlicher Detailuntersuchungen angewendet worden, in denen das Ranghäufigkeitsverhalten von Buchstaben an verschiedenen slawischen Sprachen systematisch überprüft wurde: Für das Russische in Grzybek, Kelih (2003a), Grzybek, Kelih, Altmann (2004) und Grzybek, Kelih, Altmann (2005a); für das Slowakische in Grzybek/Kelih/Altmann (2005b) und Grzybek/Kelih/Altmann (2006); für das Ukrainische in Grzybek/Kelih (2005a) und für das Slowenische in Grzybek, Kelih (2003b).

Dabei wurden vor allem in der vorherigen Forschungsgeschichte zur Diskussion gestellte Modelle auf ihre Eignung geprüft.[2] Insgesamt hat sich in diesen Untersuchungen herausgestellt, dass sich die meisten der in der Vergangenheit diskutierten Modelle als ungeeignet erweisen; lediglich ein Modell, nämlich die negative hypergeometrische Verteilung (NHG), hat sich durchgehend als geeignetes Modell für die bislang untersuchten slawischen Sprachen erwiesen .[3]

Dieses Verteilungsmodell – das im weiteren Verlauf dieser Darstellung noch detaillierter besprochen werden wird (s.u.) – weist im Vergleich mit den anderen (vgl. Fußnote 2) Verteilungen am meisten Parameter auf – und natürlich ist ein Modell um so flexibler, je mehr Parameter es aufweist. Dabei müssen bei jeder Verteilung die Parameter jeweils so bestimmt werden, dass sich das gesamte Modell möglichst gut an die jeweiligen Daten anpasst. Das geschah früher durch die Anwendung verschiedener Schätzmethoden, mit denen man sogenannte Schätzer bestimmte, die sich aus theoretischen Eigenschaften des Modells ergeben. Heutzutage werden Parameterbestimmungen entweder im Anschluss an Anfangsschätzungen oder auch ganz ohne diese mittels geeigneter Software rechnerisch bestimmt und in iterativen Prozeduren so optimiert, dass die Abweichung der theoretischen von den beobachteten Werte möglichst gering ist.[4]

Natürlich ist die Bestimmung der Parameterwerte und die damit verbundene Anpassung des Modells an die zu untersuchenden Daten im Rahmen

---

[2] Im einzelnen handelt es sich um: Zeta-Verteilung, Zipf-Mandelbrot-Verteilung, geometrische Verteilung, Good-Verteilung, Whitworth-Verteilung, negative hypergeometrische Verteilung.

[3] Interessanterweise ist sie in jüngster Zeit nicht nur für slawische Sprachen, sondern auch für das Deutsche als geeignetes Modell nachgewiesen worden (Best 2004/05, Best 2005, Grzybek 2007a,b); Details dazu müssen aus der vorliegenden Abhandlung freilich ausgeblendet werden.

[4] Aus Gründen, die an anderer Stelle ausführlich dargelegt wurden, arbeiten wir nicht mit Kurven, sondern mit diskreten Häufigkeitsmodellen. In den betreffenden Studien sind alle relevanten Ansätze, die bislang zur Modellierung von Graphemhäufigkeiten vorgeschlagen worden sind, auf ihre Eignung überprüft worden. Die Güte der Anpassungen wird dabei mit statistischen Methoden überprüft; dazu eignet sich der sog. Chiquadrat-Anpassungstest als ein Test für die Überprüfung der Güte der Anpassung. Da der Chiquadrat-Wert allerdings linear mit der Stichprobengröße zunimmt (und man insofern bei großen Stichproben – was bei Graphemhäufigkeiten eigentlich immer der Fall ist – immer schneller mit signifikanten Abweichungen konfrontiert ist), ist es sinnvoll, den Chiquadrat-Wert mit der Stichprobengröße zu relativieren und sich auf einen Diskrepanzkoeffizienten, hier $C = \chi^2/N$, zu beziehen. Dieser wird bei $C < 0.02$ als Indiz einer guten, bei $C < 0.01$ als Indiz einer sehr guten Anpassung angesehen – in diesem Fall ist somit davon auszugehen, dass die theoretische Berechnung geeignet ist, die empirisch ermittelten Werte in dem gegebenen Modell zu erfassen.

linguistischer Forschungen weder Selbstzweck und Ziel der Untersuchung; vielmehr handelt es sich hierbei um einen wesentlichen Arbeitsschritt, der im weiteren zu einer qualitativen Interpretation der quantitativen Ergebnisse führen sollte. Auf diese Art und Weise findet der entscheidende Übergang vom Auffinden und der *Beschreibung* bestimmter Regularitäten hin zu deren Interpretation bzw. *Erklärung* statt. Dieser wissenschaftstheoretisch klar definierte Schritt ist in der ausschließlich qualitativ vorgehenden Linguistik nicht selbstverständlich, was deren wissenschaftstheoretischen Status ausreichend charakterisiert. Doch auch in der quantitativen Linguistik, der letztlich gerade an Erklärungen gelegen ist, ist es bislang nur selten gelungen, eine tatsächliche Interpretation des Verhaltens von Parametern zu erarbeiten.[5]

An diesem Punkt setzt die vorliegende Untersuchung an: Ausgehend von der Beobachtung, dass zur Beschreibung von Buchstabenhäufigkeiten verschiedener slawischer Sprachen offenbar ein solch komplexes Verteilungsmodell wie die 3-parametrige NHG notwendig ist, soll ein wichtiger Schritt unternommen werden, sich einer Interpretation der Parameter dieses Modells zu nähern. Dies wäre dann als gelungen anzusehen, wenn es gelänge, das Ranghäufigkeitsverhalten nicht nur für die einzelnen Sprachen, sondern auch im Vergleich zueinander in seiner vermutlichen Systematik zu erfassen. Sollte dies zumindest ansatzweise gelingen, wäre das ein erster wichtiger Schritt, die Notwendigkeit der Komplexität dieses Modells zu verstehen.

Aus diesem Grund scheint es jedoch sinnvoll, vor den eigentlichen Analysen die bisherigen Befunde kurz zu resümieren und eingangs das Datenmaterial, an dem diese gewonnen wurden, kurz zu charakterisieren.

4. Bisherige Untersuchungen zu slawischen Sprachen

Bislang sind Graphemhäufigkeiten in vier slawischen Sprachen (Russisch, Ukrainisch, Slowakisch, Slowenisch) systematisch untersucht worden, wobei sich der Inventarumfang (I) im Bereich von $I = 25$ (Slowenisch) bis $I = 46$ (Slowakisch mit Diagraphen, s.u.) erstreckt:

(1) **Slowenisches** Textmaterial war Gegenstand einer Pilotstudie von Grzybek/Kelih (2003b): Untersucht wurden *20 Stichproben*, bei denen es sich

[5] Dabei gilt in diesem Zusammenhang das allgemeine, als „Ockham's Razor" bekannte Sparsamkeitsprinzip der Wissenschaft, dem zufolge von mehreren Theorien, die den gleichen Sachverhalt erklären, die einfachste zu bevorzugen ist, und demzufolge man in Hypothesen nicht mehr Annahmen einführen sollte, als tatsächlich benötigt werden, um einen bestimmten Sachverhalt zu beschreiben und empirisch nachprüfbare Voraussagen zu treffen. Folglich sollte man auch in Verteilungsmodelle nicht mehr Parameter einführen, als notwendig, da letztendlich alle einer Interpretation zugeführt werden müssen.

einerseits um 10 vollständige Texte, andererseits um neun Textausschnitte, -kumulationen und -kombinationen sowie ein aus diesen zusammengesetztes Korpus handelt. Im Ergebnis stellte sich heraus, dass die NHG für alle 20 Stichproben ein bestens geeignetes Modell darstellt (in allen Fällen $C < 0.02$).

(2) **Russisches** Textmaterial wurde zunächst in Grzybek, Kelih u. Altmann (2004) untersucht: analysiert wurden in dieser Studie insgesamt *38 Stichproben* unterschiedlicher Funktionalstile (sowohl vollständige *Texte* als auch *Textausschnitte*, *Textkumulierungen* und ein sich aus diesen Texten zusammensetzendes *Gesamtkorpus*). Durch diese Art der Textauswahl und -aufbereitung sollte vor allem der Faktor der Datenhomogenität kontrolliert werden. In dieser Untersuchung wurde zunächst von einem Inventarumfang von 32 Graphemen ausgegangen. Im Ergebnis stellte sich die NHG als bestens geeignetes Modell heraus: (in allen Fällen $C < 0.02$, in 33 der 38 Stichproben $C < 0.01$).- Ausgehend von diesem Befund richtete sich das Interesse in einer Anschlussuntersuchung zum Russischen (Grzybek/Kelih/Altmann 2005a) sodann auf *30 vollständige Texte*; diese wurden unter zwei verschiedenen Bedingungen alternativ analysiert: einmal mit ‚ë' und einmal ohne ‚ë' als eigenständigem Graphem – diese Versuchsbedingung ändert den Inventarumfang, der einmal $I = 32$, einmal $I = 33$ beträgt. Diese Art der Untersuchung zielte somit weniger auf eine Aussage zur sprach- und kulturpolitisch umstrittenen Funktion des ‚ë' als vielmehr auf eine detaillierte Analyse der Bedeutung des Inventarumfangs. Abgesehen davon, dass die Annahme unterschiedlicher Inventarumfänge eine (allerdings nur schwach signifikante) systematische Verschiebung der Werte von Entropie und Wiederholungsrate (repeat rate) nach sich zieht, stellte sich bezüglich eines Verteilungsmodells der Graphemhäufigkeiten unter beiden Bedingungen die NHG als gleichermaßen geeignet heraus (in 59 der 60 Stichproben $C < 0.02$).

(3) Für **slowakische** Graphemhäufigkeiten stellte sich an *30 untersuchten Texten* ebenfalls die NHG als einziges geeignetes Modell heraus (Grzybek/Kelih/Altmann 2005b und 2006); auch für das Slowakische gilt dies unter zwei verschiedenen Untersuchungsbedingungen, die sich durch den jeweils angesetzten Inventarumfang unterscheiden: dieser beträgt nämlich $I = 43$, wenn man keine Digraphen als eigene Einheiten annimmt, und $I = 46$, wenn man die Graphemkombinationen „dz", „dž" und „ch" als selbständige Grapheme zählt. Für $I = 43$ beträgt der Diskrepanzkoeffizient in 28 von 30 Fällen (zwei extrem kurze Texte fallen aus der Reihe) $C < 0.02$, davon in 10 Fällen $C < 0.01$; für $I = 46$ beläuft er sich in 25 der 30 Einzelanalysen auf einen Wert von $C < 0.02$.

(4) Weiterhin stellte sich in einer Untersuchung zu 30 **ukrainischen** Texten (Grzybek/Kelih 2005a) – der Inventarumfang liegt hier bei 33 Graphe-

men, wenn man das Apostroph nicht als eigenständiges Graphem betrachtet – die NHG als einziges valides Modell heraus. Der Diskrepanzkoeffizient liegt in allen 30 Einzelanalysen bei $C < 0.02$, davon 20 der Texte sogar $C < 0.01$.

Damit lässt sich festhalten, dass sich das Rangverhalten der Grapheme in den untersuchten Sprachen durch ein und dasselbe Modell erfassen lässt: In allen Fällen hat sich durchgängig (nur) die negative hypergeometrische Verteilung (NHG) als adäquat erwiesen. Wie auch schon andernorts dargestellt (vgl. Grzybek/Kelih/Altmann 2004), ist es für Rangierungszwecke sinnvoll, diese Verteilung um einen Schritt nach rechts zu verschieben (da es konventionell keinen Rang 0 gibt), so dass man die 1-verschobene negative hypergeometrische Verteilung (1) erhält.

(1) $$P_x = \frac{\binom{M+x-2}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}} \qquad x = 1, 2, \ldots, n+1$$

$$K > M > 0;\ n \in \{1,2,\ldots\}$$

Wie zu sehen ist, weist diese drei Parameter auf, und zwar neben dem auf den Inventarumfang zurückzuführenden Parameter n die beiden Parameter K und M. Wenn man n für eine gegebene Sprache nicht als freien Parameter definiert, sondern fixiert und $n = I$-1 ansetzt (da die 1-verschobene NHG einen Wertebereich hat, der von 1 bis n+1 geht), bleiben die beiden freien Parameter K und M, um deren Bestimmung es geht. Die Werte dieser beiden Parameter unterscheiden sich in den einzelnen Stichproben, wobei die Unterschiede auf zwei Umstände zurückzuführen sind: einerseits differieren sie für die verschiedenen Sprachen (was im Sinne eines sprachspezifischen Rangverhaltens gewertet werden kann), andererseits auch innerhalb der einzelnen Sprachen (bedingt durch eine „natürliche“ Variation der Vorkommnisse). Um das Verhalten der beiden Parameter K und M von der Tendenz her zu analysieren, liegt es somit nahe, im Hinblick auf die innersprachliche Variation für eine gegebene Sprache erstens den Mittelwert von K und M über alle Stichproben zu berechnen, und zweitens die entsprechenden, sich aus der Standardabweichung ergebenden 95%-Konfidenzintervalle. Tabelle 1 resümiert die sich aus den oben angegebenen Untersuchungen ergebenden Werte. Neben der Anzahl der untersuchten Stichproben (S) sind der jeweils zugrunde liegende Inventarumfang (I), die Mittelwerte, sowie Ober- und Untergrenze der Konfidenzintervalle für K und M angeführt.

Tab. 1: Untersuchte Sprachen und Parameterverhalten von *K* und *M*

	S	*I*	$\overline{K}$	K_u	K_o	⁂	M_u	M_o
Slowenisch_Pilot-2003	25	25	2,89	2,86	2,92	0,8115	0,8062	0,8168
Russisch_Pilot-2004	38	32	3,16	3,14	3,19	0,8186	0,8105	0,8267
Russisch-32	30	32	3,14	3,10	3,18	0,7896	0,7990	0,8202
Russisch-33	30	33	3,29	3,24	3,33	0,8227	0,8120	0,8335
Ukrainisch	30	33	2,96	2,92	3,01	0,8203	0,8082	0,8324
Slowakisch-43	30	43	4,07	4,00	4,14	0,8546	0,8389	0,8703
Slowakisch-46	30	46	4,31	4,23	4,40	0,8430	0,8276	0,8584

Auf der Basis dieser Befunde sind mittlerweile erste Versuche unternommen worden, das Parameterverhalten von *K* und *M* einer Interpretation zuzuführen (vgl. Grzybek/Kelih 2005b und Grzybek/Kelih/Altmann 2005a). Die einzelnen Argumentationsschritte, die zu dieser Interpretation geführt haben, können hier nicht im einzelnen dargelegt werden; jedenfalls konnte für Parameter *K* eine **direkte Abhängigkeit vom Inventarumfang** *I* festgestellt werden, so dass die Bestimmung von *K* de facto nur im übersprachlichen Bezug vorgenommen werden kann. Für den Parameter *M* hingegen ergab sich nur eine indirekte Abhängigkeit vom Inventarumfang *I*; wohl aber war innerhalb einer gegebenen Sprache ein **direkter Zusammenhang zwischen den Parametern *M* und *K*** festzustellen.

Abb. 1 veranschaulicht die angesprochene Tendenz der Abhängigkeit des Parameters *K* von I, wie sie sich aus den Daten der Tab. 1 ergibt; zugrunde gelegt ist jeweils der Mittelwert.

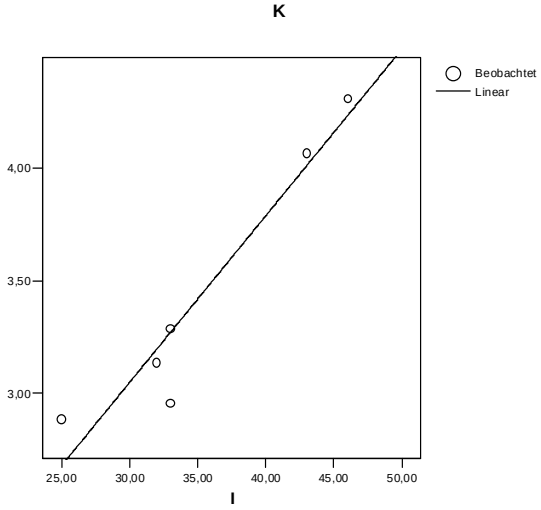


Abb. 1: Abhängigkeit des Parameters *K* vom Inventarumfang *I*

Wie zu sehen ist, bestätigt sich die oben angesprochene Tendenz: Parameter *K* korreliert hoch signifikant mit dem Inventarumfang *I* ($r = 0.96$, $p < 0.005$); diese Abhängigkeit lässt sich gut mit der linearen Gleichung $K = 0.968 \cdot I$ beschreiben.

Ebenfalls bestätigt sich an den Daten der oben angesprochene Zusammenhang zwischen den beiden Parametern *K* und *M*, wobei in den bisherigen Interpretationsansätzen – ausgehend von der obigen Interpretation des Parameters *K* – eine Abhängigkeit des Parameters *M* von *K* in Betracht gezogen wurde. Abb. 2 veranschaulicht den Zusammenhang zwischen beiden für die untersuchten Sprachen.

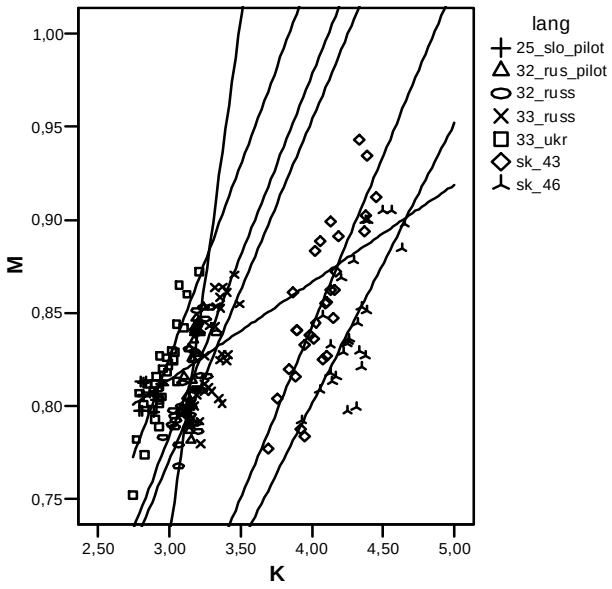


Abb. 2: Abhängigkeit des Parameters *M* von *K*

Ebenfalls bestätigt sich an den Daten der oben angesprochene Zusammenhang zwischen den beiden Parametern *K* und *M*, wobei in den bisherigen Interpretationsansätzen – ausgehend von der obigen Interpretation des Parameters *K* – eine Abhängigkeit des Parameters *M* von *K* in Betracht gezogen wurde. Abb. 2 veranschaulicht den Zusammenhang zwischen beiden für die untersuchten Sprachen.

Es zeigt sich deutlich, dass nur ein sehr schwacher übersprachlicher Zusammenhang zwischen *K* und *M* besteht, dass dieser aber innerhalb der jeweiligen Sprachen sehr ausgeprägt ist: Dabei gilt, dass je größer der Wert für *K*, desto größer auch derjenige für *M* ist.

Im Vergleich des Parameterverhaltens der verschiedenen Sprachen zeigt sich, dass dieses für die einzelnen Sprachen eine jeweils sehr ähnliche Tendenz aufweist; das spiegelt sich deutlich in dem regulären Verlauf der Regressionslinien wider, die im Bereich der Datenpunkte – d.h. in den sich für alle Sprachen ergebenden Intervallen von $4.86 \leq K \leq 2.70$ bzw. $0.74 \leq M \leq 0.94$ – kaum Überschneidungen aufweisen. Allerdings sind insgesamt bei drei der Stichproben auf unterschiedliche Art und Weise Abweichungen von diesem Gesamtbild festzustellen:

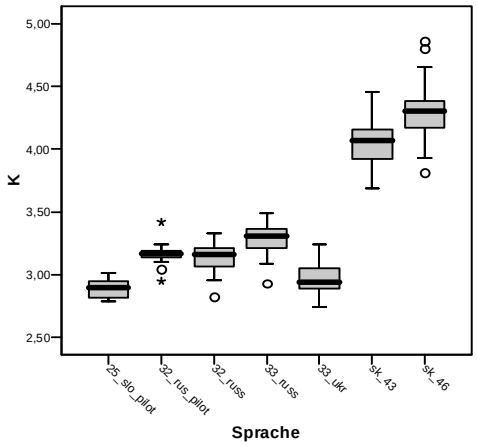
1. Die Regressionslinien für die slowakischen Stichproben mit $I = 43$ und $I = 46$ überschneiden sich und folgen nicht dem allgemeinen Trend des Parallelverlaufs;
2. Die Regressionsgerade für das Slowenische weicht ebenfalls von der Tendenz des Parallelverlaufs ab und weist Überschneidungen mit einer Reihe anderer Sprachen auf;
3. Im Falle des Ukrainischen bedingt der im Vergleich zu den anderen Sprachen vergleichsweise niedrige Parameterwert für *K*, dass die Regressionslinie sich nicht dem Inventarumfang erwartungsgemäß einordnet.

Mit diesen Beobachtungen und den sich vor dem Hintergrund dieser Befunde ergebenden Fragen können wir die „Bestandsaufnahme" bisheriger Untersuchungen abschließen; abgesehen von der deutlich herausgearbeiteten Eignung der NHG zur Modellierung der Graphemhäufigkeiten in allen untersuchten slawischen Sprachen ergibt sich im wesentlichen die Frage, ob und wie die beobachteten Abweichungen des Slowenischen und Slowakischen einerseits sowie des Ukrainischen andererseits zu erklären sind.

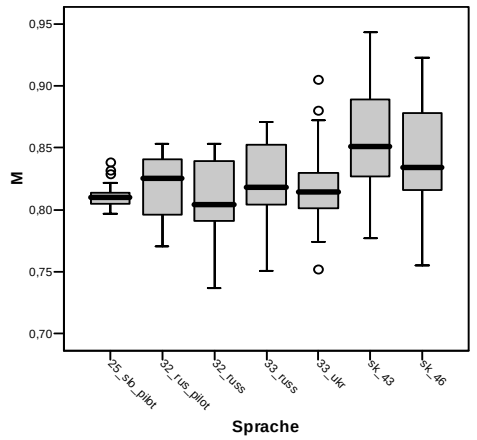
5. Stichproben, Ausreißer und Extremwerte

Ein erster Schritt, die beobachteten Abweichungen vom allgemeinen Parameterverhalten zu erklären, kann darin bestehen, allfällige Ausreißer und Extremwerte aus der Analyse zu eliminieren. Dies geschieht in der Regel durch Bezugnahme auf den sogenannten Interquartilbereich (IQR), der die mittleren 50% aller Beobachtungen umfasst. Als Ausreißer (bzw. Extremfälle) werden nun üblicherweise solche Beobachtungen bezeichnet, deren Abstand von der Ober- bzw. Untergrenze des IQR das 1.5-3fache (bzw. mehr als das 3-fache) des IQR beträgt.

Anschaulich lässt sich das in Form von sogenannten Box-Plots darstellen, bei denen Ausreißer leicht zu identifizieren sind, da sie jenseits der oberen bzw. unteren Linie liegen, welche immer durch einen Wert aus den gegebenen Daten bestimmt wird und maximal das 1,5-fache des Interquartilsabstands beträgt (wenn es keine Ausreißer gibt, handelt es sich um Minimal- bzw. Maximalwert der jeweiligen Stichprobe). Abb. 3a und 3b präsentieren die Box-Plot-Serien für die Parameterwerte von *K* und *M*.



(a) Parameter *K*



(b) Parameter *M*

Abb. 3: Boxplotserien

Deutlich zu erkennen sind die Ausreißer: Von den insgesamt 238 untersuchten Stichproben erweisen sich acht Werte als *K*-Ausreißer und sechs Werte als *M*-Ausreißer, wobei deren Anzahl in den einzelnen Stichprobenserien maximal drei Werte beträgt.

Nach Ausschluss dieser Ausreißer und einer neuerlichen Analyse des Parameterverhaltens von *K* und *M* (vgl. Abb. 4.) zeigt sich ein verändertes Bild der Regressionsgeraden:

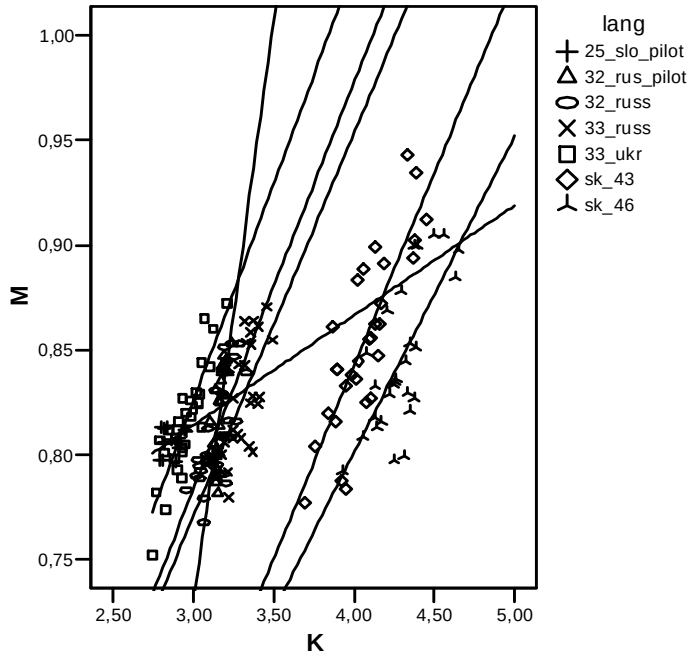


Abb. 4: Zusammenhang von *K* und *M* (nach Eliminierung von Ausreißern)

Eine erste wichtige Feststellung betrifft die Stichprobenserie Slowakisch mit $I = 46$, deren Regressionsgerade sich nunmehr nahtlos in den Paralleltrend einordnet. Eine Analyse der Ausreißerwerte zeigt, dass die drei ausgeschlossenen Stichproben mit $N = 562$, $N = 1214$ und $N = 1607$ äußerst klein sind; das lässt darauf schließen, dass sich aufgrund unzureichender Stichprobengröße die Häufigkeitsverhältnisse noch nicht ausreichend konsolidiert haben, das heißt, dass die Stichproben schlicht und einfach zu klein sind. Dies führt zur Frage der Bestimmung adäquater Stichprobengrößen, deren Behandlung freilich über den hier gegebenen Rahmen hinausgehen würde, und die es an anderer Stelle ausführlicher zu betrachten gilt (Grzybek et al. 2007).

Während damit für die Abweichung der slowakischen Stichprobenserie eine Begründung gefunden zu sein scheint, bleiben eine Reihe weiterer (bzw.

neuer) Beobachtungen erklärungsbedürftig: Denn abgesehen davon, dass die Werte für das Ukrainische nach wie vor aus dem allgemeinen Trend ausbrechen, ist auch die Tendenz für das Slowenische weiterhin abweichend, so dass hier eine andere Ursache als im Fall des Slowakischen vorliegen muss. Die Tatsache, dass auch die Stichprobenserie der russischen Pilotstudie nunmehr, nach Ausschluss der Ausreißer, ein ähnliches Bild zeigt wie die der slowenischen Pilotstudie, könnte sich dabei als Hinweis darauf werten lassen, dass das Problem seine Ursache in der spezifischen Struktur dieser Daten hat. Aufgrund der Einzeluntersuchungen kann dabei ausgeschlossen werden, dass Textsegmente, Textkumulationen oder Korpora anderen Regularitäten folgen als homogene Texte; die Ursache scheint vielmehr anders gelagert zu sein, und dem gilt es, im folgenden Abschnitt nachzugehen. Anlass für eine solche naheliegende Neu-Analyse bietet nicht zuletzt die bisherige Erfahrung mit dem Russischen: Da nämlich der allgemeine Trend für das Russische mittlerweile durch systematische Untersuchungen, die im Anschluss an die Pilotstudie durchgeführt worden, abgesichert ist und sich die Abweichung nur in den Daten der Pilotstudie äußert, scheint es sinnvoll, auch für das Slowenische eine neue, systematische Untersuchung durchzuführen. Dies scheint um so mehr gerechtfertigt, als das Slowenische innerhalb der slawischen Sprachen bei einer Inventargröße von $I = 25$ den geringsten Inventarumfang aufweist.

6. Slowenische Graphemhäufigkeiten: Die Notwendigkeit systematischer Untersuchungen

De facto wurden Graphemhäufigkeiten und Möglichkeiten ihrer theoretischen Modellierung im Slowenischen bislang nie systematisch untersucht. Das betrifft nicht nur die oben genannte Pilotstudie und die Qualität der darin verwendeten Daten, sondern gilt für die Geschichte der Erforschung slowenischer Graphemhäufigkeiten schlechthin: Ohne einen Anspruch auf Vollständigkeit erheben zu wollen, ist davon auszugehen, dass die Frage von Graphemhäufigkeiten im Slowenischen – wie quantitativ-linguistische Untersuchungen überhaupt – nur fragmentarisch und in der Regel ohne weiteren theoretischen Hintergrund verfolgt worden sind. Erste umfassendere quantitative Untersuchungen des Slowenischen gehen auf Poniž (1974) zurück, der bezogen auf Buchstaben erste deskriptive Statistiken zur Häufigkeit von Buchstaben (zum Teil getrennt nach Groß- und Kleinbuchstaben, Wortanfang, Wortmitte) usw. vorgelegt hat. Neuere Untersuchungen zum slowenischen Buchstabenbestand finden sich in weiterer Folge bei Jakopin (1995), wo die Buchstabenhäufigkeit des *Slovar Slovenskega Knjižnega Jezika* zu finden ist. Eine Buchstabenstatistik gibt es auch für das *Rückläufige Wörterbuch des*

Slowenischen (vgl. Hajnšek-Holz/Jakopin 1996: 701ff.). Schließlich finden sich Buchstabenstatistiken auch in Jakopin (2002), wo neben den einzelnen Häufigkeiten von Buchstaben auch einfache Entropie-Berechnungen angeführt sind, die allerdings nicht weiter interpretiert werden.

Insofern stellt die oben erwähnte Pilotstudie zum Slowenischen einen qualitativen Neubeginn dar, der über die empirische Erhebung von Graphemhäufigkeiten hinausgehend auf theoretische Fragen der Sprachmodellierung abzielt. Dennoch ist – wie sich oben herausgestellt hat – das Material dieser Pilotstudie nicht dazu geeignet, als Grundlage einer systematischen Untersuchung zu dienen; es unterscheidet sich von der Qualität der anderen von uns untersuchten Sprachen in mehrerer Hinsicht:

1. Es wurden nur 20 (nicht, wie in den anderen Sprachen, 30) Stichproben untersucht;
2. nur bei neun der 20 Stichproben handelte es sich um vollständige Texte;
3. bei den 11 übrigen Stichproben handelte es sich um spezifische Teilmengen der neun vollständigen Texte;
4. es wurden fast ausschließlich Stichproben literarischer Texte verwendet, wobei freilich ein systematischer Einfluss von Textsorten auf Graphemhäufigkeiten bislang nie kontrolliert untersucht worden ist.

Während die beiden ersten Punkte nicht unbedingt wesentlich ins Gewicht fallen, ist der dritte von größerer Bedeutung, da durch diesen Umstand bedingt ist, dass sich die Datenstruktur der neun Basistexte im Grunde genommen dupliziert; das gilt auch für das Gesamtkorpus, das sich aus den anderen Stichproben zusammensetzte. Aus diesem Grunde zeichnet sich das Datenmaterial nicht durch die nötige Streuungsbreite aus, wie es in den anderen Stichprobenserien gewährleistet ist.[6]

Aus den genannten Gründen soll im folgenden eine mit den anderen Sprachen vergleichbare Datenbasis untersucht werden, die dann auch einen Vergleich mit diesen erlaubt. Tabelle 2 stellt das Textmaterial in übersichtlicher Form dar. Bei den Texten handelt es sich um einzelne Roman-Texte, journalistische Kommentare, wissenschaftliche Texte (Kapitel aus Diplomarbeiten und Artikel), Privatbriefe und Predigten. In diesen Texten wird die Häufigkeit der einzelnen Buchstaben untersucht, wobei von einem Grapheminventar von $I = 25$ ausgegangen wird: A,B,C,Č,D,E,F,G,H,I,J,K,L,M,N,O,P,R,S,Š, T,U,V,Z,Ž.

In Tabelle 2 ist mit *N* der Gesamtumfang in der Anzahl von Graphemen pro Text bezeichnet.

[6] Lediglich die Pilotstudie zum Russischen gleicht derjenigen zum Slowenischen in eben dieser Hinsicht, was auch die abweichende Tendenz dieser Daten vom Gesamtbild erklären würde.

Tab. 2: Text- und Datenbasis (Slowenisch)[7]

Nr.	Autor/Zeitschrift	Text	*N*	Nr.	Autor/Zeitschrift	Text	*N*
1	div.	Diplomarbeit 1	39176	16	Josip Jurčič	Privatbrief 1	632
2		Diplomarbeit 2	61919	17		Privatbrief 2	1027
3		Diplomarbeit 3	10057	18		Privatbrief 3	1374
4		Diplomarbeit 4	8505	19		Privatbrief 4	3637
5		Diplomarbeit 5	69251	20		Privatbrief 5	2011
6		*Kommentar 1*	2871	21		*Hlapec Jernej […] Kapitel 2*	4027
7		*Kommentar 2*	3489	22		*Kapitel 3*	4288
8		*Kommentar 3*	1900	23		*Kapitel 4*	3381
9		*Kommentar 4*	3889	24		*Kapitel 5*	3397
10	Delo	*Kommentar 5*	4237	25	Ivan Cankar	*Kapitel 7*	4191
11	div.	Predigt 1	4606	26	div.	wiss. Artikel 1	2006
12		Predigt 2	3324	27		wiss. Artikel 2	6277
13		Predigt 3	2310	28		wiss. Artikel 3	27146
14		Predigt 4	3266	29		wiss. Artikel 4	3617
15		Predigt 5	5402	30		wiss. Artikel 5	22522
					Gesamt		**313735**

Für jeden Text wird in einem ersten Schritt die Vorkommenshäufigkeit der einzelnen Buchstaben bestimmt; diese Häufigkeiten werden sodann in eine Ranghäufigkeitsverteilung transformiert.

Dieses Vorgehen lässt sich exemplarisch am Gesamtkorpus dieser Texte veranschaulichen: Fügt man die 30 einzelnen Texte zu einem Gesamtkorpus zusammen, so beläuft sich der Umfang dieses Korpus auf $N = 313735$ slowenische Grapheme. Tab. 2 gibt die absoluten und relativen Häufigkeiten für die 25 Grapheme an.

[7] Die Texte sind der im Rahmen des FWF-Projekts P-15485 »Wortlängenhäufigkeiten in Texten slawischer Sprachen« eingerichteten Text-Datenbank entnommen (vgl.: http://www-gewi.uni-graz.at/quanta).

Tab. 3: Graphemhäufigkeiten im Gesamtkorpus

Graphem	f(i)	$f_{rel}(i)$	Graphem	f(i)	$f_{rel}(i)$
a	31891	0,10	**m**	9568	0,03
b	4608	0,01	**n**	22905	0,07
c	2463	0,01	**o**	31122	0,10
č	5361	0,02	**p**	10514	0,03
d	10216	0,03	**r**	16084	0,05
e	32036	0,10	**s**	14668	0,05
f	497	0,00	**š**	2606	0,01
g	5055	0,02	**t**	16088	0,05
h	2554	0,01	**u**	7446	0,02
i	27150	0,09	**v**	15221	0,05
j	14043	0,04	**z**	6413	0,02
k	10517	0,03	**ž**	1675	0,01
l	13034	0,04			

Ordnet man die Vorkommenshäufigkeiten der einzelnen Grapheme in absteigender Reihenfolge, so ergibt sich die Ranghäufigkeitstabelle, die für das Gesamtkorpus in Abb. 5 zu sehen ist.

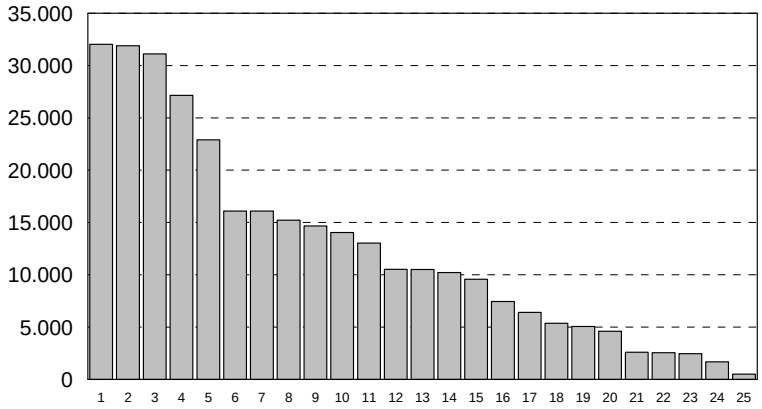


Abb. 5: Ranghäufigkeitsverteilung (Korpus slowenischer Grapheme, *N* = 313735)

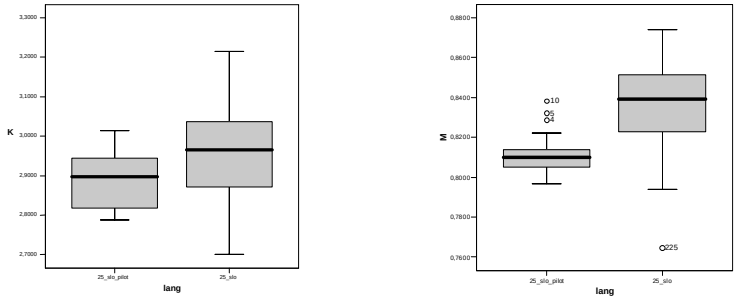
Im weiteren geht es um die theoretische Modellierung der Ranghäufigkeiten in den 30 slowenischen Texten, wobei wir uns unter Bezugnahme auf die einleitende Darstellung auf die NHG und die Bestimmung der Parameterwerte beschränken. Tabelle 4 enthält die Ergebnisse der Anpassungen; folgende Informationen sind enthalten: Textnummer, die sich aus der Anpassung der Verteilungsmodelle ergebenden Parameterwerte (K und M), der X^2-Wert (die Anzahl der Freiheitsgrade beträgt konstant $FG = 21$) sowie der Wert des Diskrepanzkoeffizienten C. Es ergeben sich in der Zusammenschau für das Slowenische insgesamt gute Ergebnisse: Der Diskrepanzkoeffizient liegt für alle 30 Stichproben im Intervall von $0.022 \geq C \geq 0.0055$; in 26 der 30 Einzelanalysen ist $C < 0.02$, davon in sechs Fällen $C < 0.01$. Damit bestätigt sich, dass die NHG insgesamt ein durchgehend gutes Modell ist, um die Graphemhäufigkeiten im Slowenischen zu erfassen.

Tab. 4: Ergebnisse der Anpassung der NHG (30 slowenische Texte)

Nr.	*K*	*M*	X^2	C	Nr.	*K*	*M*	X^2	C
1	3,0970	0,8665	442,69	0,0113	**16**	2,9786	0,8493	10,24	0,0162
2	3,1663	0,8608	557,27	0,0090	**17**	2,7001	0,7648	17,87	0,0174
3	2,9688	0,8402	96,55	0,0096	**18**	2,8715	0,8283	18,14	0,0132
4	2,9965	0,8439	86,75	0,0102	**19**	2,9445	0,8384	79,65	0,0219
5	3,0199	0,8521	747,91	0,0108	**20**	2,9801	0,8423	39,82	0,0198
6	3,0610	0,8581	44,50	0,0155	**21**	3,0364	0,8403	28,59	0,0071
7	2,9797	0,8315	42,22	0,0121	**22**	2,8438	0,7938	46,31	0,0108
8	2,9484	0,8383	29,64	0,0156	**23**	2,8435	0,7964	23,67	0,0070
9	3,0614	0,8550	52,11	0,0134	**24**	2,8619	0,8268	40,09	0,0118
10	2,8028	0,8134	71,61	0,0169	**25**	2,9670	0,8310	54,48	0,0130
11	2,9091	0,8353	25,33	0,0055	**26**	2,9235	0,8168	36,11	0,0180
12	3,0931	0,8515	62,82	0,0189	**27**	3,0416	0,8512	80,35	0,0128
13	3,2140	0,8740	28,41	0,0123	**28**	2,8787	0,8228	336,61	0,0124
14	2,8899	0,8426	31,35	0,0096	**29**	2,9637	0,8593	79,57	0,0220
15	2,7831	0,8166	105,34	0,0195	**30**	2,8668	0,8121	403,14	0,0179

Wie den in Tab. 4 dargestellten Ergebnissen zu entnehmen ist, sind die Werte für K und M durchaus unterschiedlich im Vergleich zu denen der slowenischen Pilotstudie: Bei einem Mittelwert von $K = 2.96$ liegen Ober- und Untergrenze des Konfidenzintervalls bei $K_u = 2.91$ und $K_o = 3.00$, und bei

einem Mittelwert von $M = 0.8351$ betragen Ober- und Untergrenze des Konfidenzintervalls $M_u = 0.8263$ und $M_o = 0.8439$. Deutlich zu erkennen sind die Unterschiede der Datenstruktur auch an den in Abb. 6a und 6b dargestellten Box-Plots: Wie zu sehen ist, deckt die systematische Stichprobenserie im Vergleich zur Pilotstudie einen breiteren Bereich ab und weist nicht zuletzt deshalb auch keine Ausreißer auf.



(a) Parameter *K* **(b)** Parameter M

Abb. 6: Boxplots für die Parameter

Es ist offensichtlich, dass diese Befunde eine Erklärung für die oben beobachtete Abweichung vom Parallelitätstrend darstellen. Insofern liegt es nahe, die slowenischen ebenso wie die russischen Daten der Pilotstudien durch diejenigen der systematischen Studien zu ersetzen und in das Gesamtschema der Abhängigkeiten für alle Sprachen mit den definierten Inventarumfängen einzusetzen. Abb. 7 zeigt das sich nunmehr ergebende Gesamtbild:

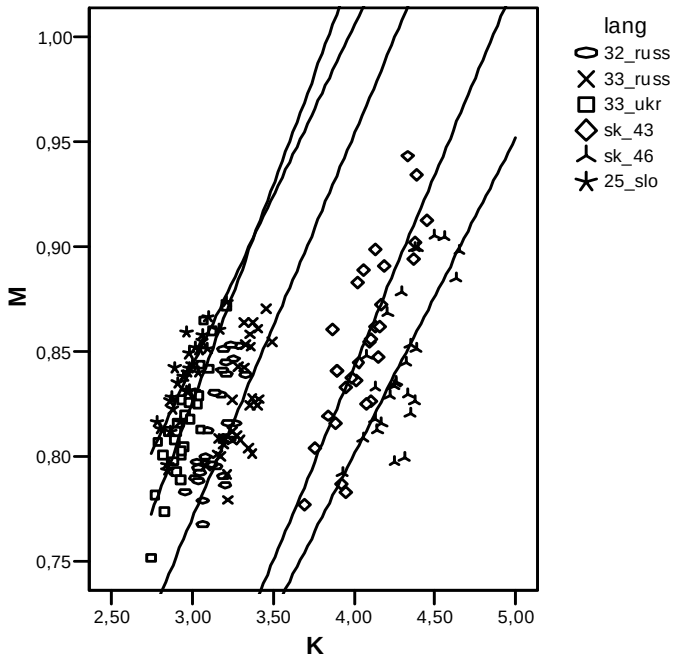


Abb. 7: Zusammenhang von *K* und *M* für alle untersuchten Sprachen (nach Elimination von Ausreißern)

Wie zu sehen ist, ähnelt der Verlauf der Regressionsgeraden für die slowenischen Daten nunmehr demjenigen aller anderen Sprachen, auch wenn sie sich nicht, wie eindeutig zu sehen ist, im Bereich der beobachteten Datenpunkte parallel zu diesen verhält. Tab. 5 enthält die Regressionsgleichungen für alle Sprachen (nach Elimination der Ausreißer). Die Regressionsgeraden folgen der Gleichung $y = b + ax$ (in unserem Fall also $M = b + aK$). Hierbei ist b eine Konstante, welche die Höhe der Regressionsgeraden (Schnittpunkt mit der y-Achse) angibt, und a ist der Regressionskoeffizient, der die Steilheit des Anstiegs bzw. Abfalls der Geraden bestimmt.

Tab. 5: Regressionskoeffizienten: $M_i = b_i + a_i \cdot K_i$

	I	*A*	*b*	R^2
Slowenisch-25	25	,161982	,357218	,73537
Russisch-32	32	,194101	,201173	,51986
Russisch-33	33	,182660	,222725	,48868
Ukrainisch	33	,206958	,204891	,76003
Slowakisch-43	43	,183903	,106713	,68662
Slowakisch-46	46	,150906	,197374	,52235

Es liegt nahe, zumindest die Regressionsgerade des Slowenischen mit der benachbarten des Ukrainischen zu vergleichen und den Unterschied zwischen beiden auf Signifikanz zu testen. Dies geschieht bei linearen Zusammenhängen über die *t*-verteilte Prüfgröße

(2) $$t = \frac{|a_1 - a_2|}{\sqrt{\frac{s^2_{y1.x1} \cdot (n_1 - 2) + s^2_{y2.x2} \cdot (n_2 - 2)}{n_1 + n_2 - 4} \cdot \left(\frac{1}{Q_{x1}} + \frac{1}{Q_{x2}} \right)}}$$

bei $FG = n_1 + n_2 - 4$ Freiheitsgraden mit $Q_x = \sum (x - \overline{x})^2$.

Im Ergebnis weist der Vergleich der beiden Regressionskoeffizienten den Unterschied zwischen dem Slowenischen und dem Ukrainischen bei einem Wert von $t = 1.52$ und $FG = 53$ als nicht signifikant aus ($p = 0.13$).[8]

Mit diesem Ergebnis stellen sich zusammenfassend und abschließend zwei Fragen, einmal in Richtung auf ein den untersuchten Sprachen gemeinsam zugrunde liegendes Parameterverhalten, einmal im Hinblick auf sprachspezifische Unterscheidungsmöglichkeiten; beide Fragerichtungen widersprechen sich nicht prinzipiell, auch wenn sie in unterschiedliche Richtung zielen:

1. Inwiefern folgen die einzelnen Sprachen einem einheitlichen Trend?
2. Inwiefern lassen sich aufgrund der Parameterwerte für *K* und *M* die Sprachen voneinander trennen?

[8] Bezeichnenderweise ist der Unterschied der beiden Regressionskoeffizienten für die slowenischen Daten der Pilotstudie – mit der Regressionsgleichung $M = 0.6569 + 0.0524\ K$ – und der systematischen Studie bei einem Wert von $t = 3.31$ und $FG = 41$ hoch signifikant ($p < 0.005$).

7. Sprachübergreifende Trends: Ein allgemeines Regressionsmodell

Im Hinblick auf eine sich möglicherweise übersprachlich manifestierende Einheitlichkeit des Trends lässt sich die Frage dahingehend exakter formulieren, als sich spezifischer untersuchen lässt, ob die auf der Abhängigkeit des Parameters *M* von *K* basierenden Regressionsgeraden insgesamt einen einheitlichen Trend aufweisen. Dies mündet in die Frage nach der Signifikanz von Unterschieden zwischen den Regressionskoeffizienten und der Parallelität der Regressionsgeraden.

Ein geeignetes Verfahren zur Überprüfung ist der multiple partielle *F*-Test; dieser wird üblicherweise im Zusammenhang mit multiplen linearen Regressionen angewendet (s. Kleinbaum et al 1998). Dabei geht es um die Frage des zusätzlichen Beitrags von unabhängigen Variablen, hinausgehend über im Modell bereits vorhandene Variablen. Der *F*-Test richtet sich dabei auf die Frage der Erweiterung eines bestehenden Modells durch das gleichzeitige Hinzufügen von zwei oder mehreren Variablen. Ein solches multiples Modell hat die folgende vollständige Form:

(3) $Y = \alpha + \beta_1 X_1 + ... + \beta_p X_p + \beta_1^* X_1^* + ... + \beta_k^* X_k^* + \varepsilon$.

Dabei ist *Y* die abhängige Variable, α die Regressionskonstante, ε ein Zufallsfehler; *X*i und X_i^* sind die unabhängigen Variablen, β_i und β_i^* die Regressionskoeffizienten. Die zu testende Nullhypothese (H_0) besagt, dass $X_1^*, X_2^*, ..., X_k^*$ nicht signifikant zur Vorhersage von *Y* beitragen, wenn X_1, X_2, …, X_k bereits im Modell enthalten sind, bzw. dass für das volle Modell gilt: $H_0 : \beta_1^* = \beta_2^* = ... = \beta_k^* = 0$. Aus dieser (zweiten) Formulierung geht die reduzierte Form des Modells hervor:

(4) $Y = \alpha + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon$.

Im Prinzip wird also die durch die Hinzufügung von X_i^* bedingte zusätzliche Summe der Abweichungsquadrate (*SAQ*) berechnet, mit der sich die folgende *F*-Statistik durchführen lässt:

(5) $$F = \frac{\left[SAQ_{reg}(vollständiges\ Modell) - SAQ_{reg}(reduziertes\ Modell) \right] / k}{SAQ_{res}(vollständiges\ Modell) / (n - p - k - 1)}$$

Dabei bezeichnet SAQ_{reg} die Summe der Abweichungsquadrate des (vollständigen bzw. reduzierten) Regressionsmodells und SAQ_{res} den Mittelwert der

Quadratsumme der Residuen des vollständigen Modells; *n* entspricht der Stichprobengröße, *p* der Anzahl der Regressionskoeffizienten des reduzierten Modells, und *k* der Anzahl der Regressionskoeffizienten, die unter Annahme der Nullhypothese H_0 gleich Null sind.

In unserem Fall besteht die Gesamtstichprobe von $n = 172$ „Texten" aus sechs Teilstichproben, die jeweils einer Sprache (mit gegebenenfalls alternativ festgelegtem Inventarumfang) zugeordnet werden. Das es sich bei dieser Zuordnung um eine nominale Kategorie handelt, und kategoriale Prädiktoren nicht direkt in ein Regressionsmodell eingeführt und sinnvoll interpretiert werden können, muss die enthaltene Information anders kodiert (re-kodiert) werden. Dazu nimmt man in der Regel eine sogenannte Dummy-Codierung vor. Hierbei werden Variablen in Untervariablen, sog. Indikatoren, zerlegt und dichotom kodiert; jede Merkmalsausprägung wird dabei gesondert nach „vorhanden" (1) bzw. „nicht vorhanden" (0) beurteilt. Die Zugehörigkeit von Fällen zu verschiedenen Stichproben lässt sich so auch als Variable mit den Codierungen 0 und 1 betrachten. Der Vorteil einer solchen Null-Eins-Codierung liegt vor allem darin, dass – unabhängig vom ursprünglichen Skalenniveau – Dummy-Variablen statistisch wie intervallskalierte Variablen behandelt werden können.

Eine kategoriale Variable mit *k+1* Ausprägungen wird also in *k* Variablen mit jeweils zwei Ausprägungen transformiert; wenn unsere Variable „Sprache mit gegebenem Inventarumfang" also ursprünglich sechs Ausprägungen hat, dann lassen sich fünf dichotome Variablen (D_1 bis D_5) konstruieren, die dieselbe Information enthalten wie die eine kategoriale. In unserem Fall erhalten wir somit das in Tab. 6 dargestellte Schema:

Tab. 6: Schema der Kodierung und Dummy-Codierung

		D_1	D_2	D_3	D_4	D_5
Slowenisch_25	*8*	0	0	0	0	0
Russisch_32	*3*	1	0	0	0	0
Russisch_33	*4*	0	1	0	0	0
Ukrainisch_33	*5*	0	0	1	0	0
Slowakisch_43	*6*	0	0	0	1	0
Slowakisch_46	*7*	0	0	0	0	1

Das beschriebene Verfahren lässt sich anwenden auf den Spezialfall der Frage nach der Parallelität der Regressionsgeraden, da eine Regressionsgerade dann als parallel zu einer bzw. mehreren anderen angesehen werden kann,

wenn sich der Vorhersagewert von *Y* durch die Hinzufügung der zusätzlichen Variablen nicht signifikant ändert. Das reduzierte Modell für M ?? lautet dann:

(6) $M = \alpha + \beta_1 \cdot K + \beta_2 \cdot D_1 + \beta_3 \cdot D_2 + ... + \beta_6 \cdot D_5 + \varepsilon$

Das bedeutet, dass die Regressionsgeraden der sechs Gruppen alle parallel sind mit gleicher Steigung β_1 ($p = 6$). Damit sind die Voraussetzungen gegeben, auf der Basis der vorgenommenen Dummy-Codierung der Variablen ‚Sprache mit gegebenem Inventarumfang' einen Vergleich zwischen den oben als ‚vollständig' und ‚reduziert' bezeichneten Modellen durchzuführen, indem man zu den Dummy-Variablen X_2 = D_1 … X_6 = D_5 die jeweiligen Produkte von X_i mit *K* als Variablen $X_1^*, X_2^*, ..., X_k^*$ hinzufügt. So erhält man in unserem Fall für das vollständige Modell 11 Parameter (*K*, fünf dummy-codierte Variablen, sowie fünf Dummyprodukte), für das sich die Werte von $SAQ_{reg} = 0.135$ sowie $SAQ_{res} = 0.057$ ergeben. Im Vergleich dazu erhält man für das reduzierte Modell, welches neben *K* fünf dummy-codierte und damit insgesamt sechs Variablen enthält, Werte von $SAQ_{reg} = 0.134$; die Summe der Abweichungsquadrate der Residuen des reduzierten Modells beträgt $SAQ_{res} = 0.058$.

Die für die Berechnung des *F*-Wertes notwendige Anzahl von *k* (siehe Formel 5) ergibt sich durch die Differenz der Variablen des vollständigen Modells (11) und des reduzierten Modells ($p = 6$), so dass in unserem Fall $k = 5$, was gleichzeitig der Anzahl der Zählerfreiheitsgrade entspricht. Der ebenfalls erforderliche Mittelwert der Quadratsumme der Residuen ergibt sich aus dem Quotienten der Quadratsumme der Residuen (SAQ_{res}) und der Anzahl der Nennerfreiheitsgrade, die sich berechnen als $m = n - p - k - 1$, so dass in unserem Fall $m = 172 - 11 - 1 = 160$. Damit lässt sich der *F*-Wert berechnen als

$$F_{(FG_1=5, FG_2=160)} = \frac{(0.0135 - 0.0134)/5}{0.058/160} = 0.056$$

Wie sich in entsprechenden Tabellen nachschauen oder leicht berechnen lässt, entspricht dieser *F*-Wert einer Wahrscheinlichkeit von $p = 0.73$, was weit von jeglicher Signifikanzgrenze entfernt ist, so dass die Nullhypothese (H_0: $\beta_1^* = \beta_2^* = ... = \beta_5^* = 0$), der zufolge die Regressionsgeraden parallel verlaufen, beizubehalten ist.

Damit lässt sich zusammenfassend festhalten, dass die Abhängigkeiten der Parameter *K* und *M* der NHG sich in der Tat übersprachlich identisch verhalten und einem gemeinsamen Regressionsmodell folgen. Diesem Modell zufolge ergibt sich ein gemeinsamer Regressionskoeffizient von $\hat{b} \approx 0.18$,

so dass sich der Parameter *M* der NHG für die untersuchten Sprachen als $M \approx \hat{a} + 0.18 \cdot K$ abschätzen ließe. Die Unterschiede zwischen den Sprachen ergeben sich durch die abweichenden Achsenabschnitte (Intercepts).

Aus dem beschriebenen allgemeinen Modell lassen sich nun die einzelnen Gruppen (Sprachen mit gegebenen Inventarumfang) als Spezialfälle ableiten. Aufgrund des Befundes, dass die Nullhypothese H_0 beizubehalten ist, reicht es, hierfür das reduzierte Modell heranzuziehen; es ergibt sich (unter Auslassung des Schätzfehlers ε):

Gruppe 1: $\alpha + \beta_1 \cdot K$

Gruppe 2: $(\alpha + \beta_2) + \beta_1 \cdot K$

Gruppe 3: $(\alpha + \beta_3) + \beta_1 \cdot K$

Gruppe 4: $(\alpha + \beta_4) + \beta_1 \cdot K$

Gruppe 5: $(\alpha + \beta_5) + \beta_1 \cdot K$

Gruppe 6: $(\alpha + \beta_6) + \beta_1 \cdot K$

Für unsere sechs Sprachengruppen mit den jeweiligen Inventargruppen ergeben sich somit insgesamt die folgenden speziellen Regressionsmodelle:

Slowenisch_25 $M = 0.085 + 0.18K$
Russisch_32 $M = 0.028 + 0.18K$
Russisch_33 $M = 0.014 + 0.18K$
Ukrainisch_33 $M = 0.066 + 0.18K$
Slowakisch_43 $M = 0.18K - 0.092$
Slowakisch_46 $M = 0.18K - 0.142$

Überführt man die Intercepts der einzelnen Sprachen in ein Regressionsmodell mit dem Inventarumfang *I* als unabhängiger Variable, ergibt sich eine hoch signifikante Korrelation ($r = 0.96$, $p < 0.001$). Abb. 8 veranschaulicht den linearen Zusammenhang:

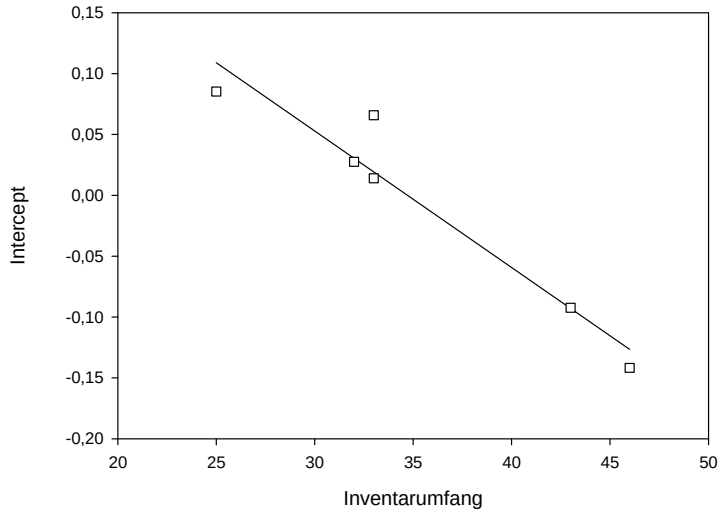


Abb. 8: Zusammenhang zwischen Intercepts des Regressionsmodells und Inventarumfang *I*

Mit diesem Befund ist im Prinzip die anfangs als Untersuchungsziel deklarierte Systematik des Parameterverhaltens in einzel- und übersprachlicher Hinsicht geleistet. Es bleibt abschließend im Rahmen der hier vorgestellten Überlegungen lediglich offen, inwiefern sich die einzelnen Sprachen mit den gegebenen Inventarumfängen aufgrund der Parameterwerte voneinander differenzieren lassen.

8. Sprachspezifische Charakteristika: Diskriminanzanalysen

Wenn aufgrund der obigen Befunde davon auszugehen ist, dass die einzelnen untersuchten Sprachen eine gemeinsame Tendenz aufweisen, die durch ein ihnen gemeinsames, einheitliches Modell beschrieben werden kann, stellt sich die zweite oben angesprochene Frage, inwiefern sich die Sprachen dennoch aufgrund der einzelsprachlich zu differenzierenden Parameterwerte trennen lassen.

Zur Untersuchung dieser Frage eignen sich am besten sog. *multivariate Diskriminanzanalysen*: Hier werden die einzelnen Fälle (Texte) zunächst bestimmten Gruppen zugeordnet (im gegebenen Fall den verschiedenen Sprachen). In der Analyse werden dann die einzelnen Fälle auf der Basis von spezifischen Prädiktorvariablen (wie etwa die Parameterwerte für *K* und *M*)

Gruppen zugeordnet; dabei werden die Variablen spezifischen (in der Regel linearen) Transformationen unterzogen, um so eine optimale Diskrimination der einzelnen Fälle zu erreichen. Auf diese Art und Weise kann getestet werden, inwiefern die qualitative apriori-Zuordnung dem quantitativen Informationsbestand der untersuchten Texte entspricht. Im Ergebnis lässt sich dann sagen, wie viele der Fälle (bzw. wie viel Prozent der Fälle) aufgrund der verwendeten Kenngrößen den a priori zugeordneten Kategorien „richtig" zugeordnet werden, was als Indiz für die Güte der gewählten Prädiktorvariable(n) gewertet werden kann.

Die Frage, inwiefern das beschriebene Modell sich im Vergleich der einzelnen Sprachen unterschiedlich ausnimmt, beinhaltet somit nicht Kenngrößen der beobachten Verteilung (wie z.B. deren Mittelwert, Schiefe, oder anderes), sondern theoretische Größe des Verteilungsmodells. Dazu stehen uns zunächst die beiden Parameter *K* und *M* zur Verfügung. Abb. 9 veranschaulicht das Ergebnis der Diskriminanzanalysen mit *K* und *M* als den beiden Prädiktorvariablen.

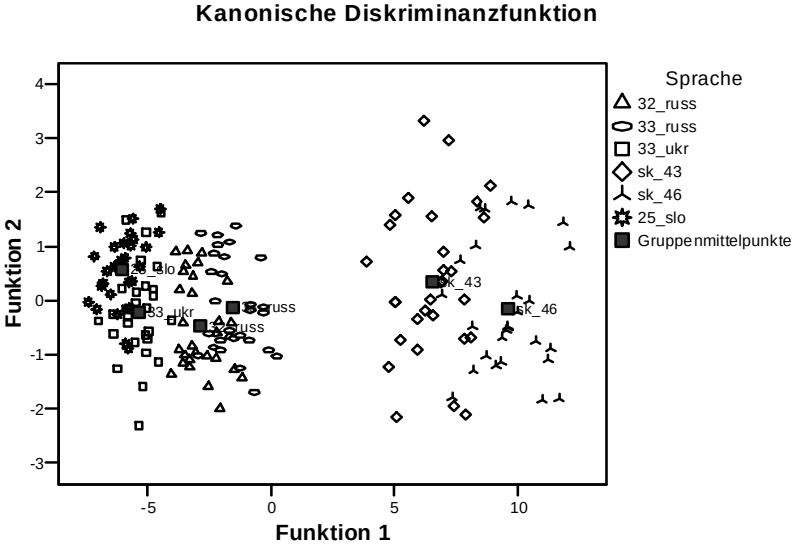


Abb. 9: Diskriminanzanalysen mit *K* und *M*

Wie zu sehen ist, erweisen sich mit den Prädiktorvariablen *K* und *M* insgesamt 82.6% der Zuordnungen als korrekt – ein Ergebnis, das nicht schlecht ist, aber auch nicht wirklich überzeugt und eher dafür spricht, dass sich mit

den Parameterwerten des Verteilungsmodells keine gute Diskrimination erreichen lässt.[9]

Damit liegt eine Situation vor, die sich wie folgt umreißen lässt: Auf der einen Seite gibt es offensichtliche Beziehungen zwischen der Inventargröße *I* und den Parametern *K* und *M*, deren konkrete Werte die Tendenz einer einzelsprachlichen Differenzierung aufweisen, die sich mit den beiden Parametern *K* und *M* allerdings nicht befriedigend erreichen lässt. Dies spricht dafür, unter Berücksichtigung dieser genannten Faktoren eine weitere Kenngröße des theoretischen Modells zu bestimmen, die sich als (weiterer) geeigneter Diskriminanzfaktor erweisen könnte.

Es liegt nahe, hierbei auf den theoretischen Mittelwert der Verteilung Bezug zu nehmen, der für die negative hypergeometrische Verteilung genau über die drei oben genannten Größen *I*, *K* und *M* definiert ist als

(7) $$m_1 = \frac{M \cdot I}{K}$$

bzw., für die um 1 verschobene Verteilung, als

(8) $$m_1 = \frac{M \cdot I}{K} + 1.$$

Abb. 10a zeigt zunächst die Beziehung zwischen den empirisch beobachteten Mittelwerten der 172 Stichproben (nach Elimination der Ausreißer) und den sich für diese aus der obigen Formel ergebenden theoretischen Mittelwerten des Verteilungsmodells; wie zu sehen ist, erweist sich der Zusammenhang als nahezu perfekt ($r = .997$, $p < 0.001$).

[9] Mit nur einem der beiden Parameter sind die Ergebnisse natürlich noch schlechter: Nur mit *K* beträgt das Ergebnis 62.2%, nur mit *M* sogar nur 29.1% korrekter Zuordnungen.

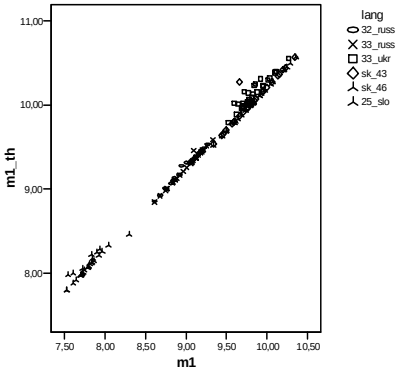


Abb. 10a: Zusammenhang von empirischen und theoretischen Mittelwerten

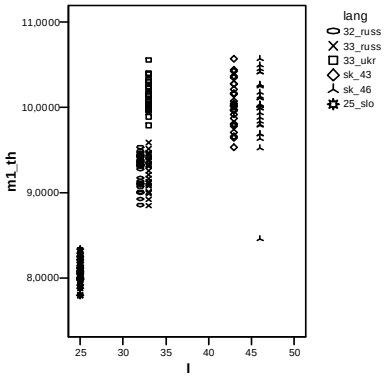


Abb. 10b: Theoretische Mittelwerte in Abhängigkeit vom Inventarumfang

Abb. 10b bildet den (theoretischen) Mittelwert in Abhängigkeit vom Inventarumfang ab; deutlich erkennbar ist eine allgemeine Tendenz, der zufolge der Mittelwert mit zunehmendem Inventarumfang ansteigt. Interessanterweise sind allerdings die Unterschiede zwischen den beiden russischen Stichprobenserien (mit $I = 32$ bzw. $I = 33$) und den beiden slowakischen (mit $I = 43$) bzw. $I = 46$) jeweils nicht signifikant. Eine klare Ausnahme stellt das Ukrainische dar, was sich mit den obigen Befunden deckt, nunmehr aber auf einen offensichtlichen Zusammenhang zwischen den Parametern K und M und dem Mittelwert der Verteilung hinweist.

Schauen wir uns vor diesem Hintergrund die Ergebnisse der Diskriminanzanalysen unter Berücksichtigung des (theoretischen) Mittelwertes als Prädiktorvariable an. Tab. 7 fasst die Ergebnisse zusammen, wobei das Symbol ✓ die jeweils in die Diskriminanzanalyse aufgenommene(n) Variable(n) kennzeichnet.

Tab. 7: Verwendete Prädikatorenvariablen und prozentuelle Trennung

K	*M*	m_1	%
✓			62,20
	✓		29,10
		✓	51,70
✓	✓		82,60
✓		✓	82,60
	✓	✓	61,00
✓	✓	✓	99,40

Es zeigt sich deutlich, dass (nur) eine Aufnahme aller drei genannten Variablen – *K*, *M*, und m_1 – ein überzeugendes Ergebnis liefert, das mit mehr als 99% korrekter Zuordnungen in der Tat überwältigend ist. Abb. 11 veranschaulicht dieses Ergebnis.

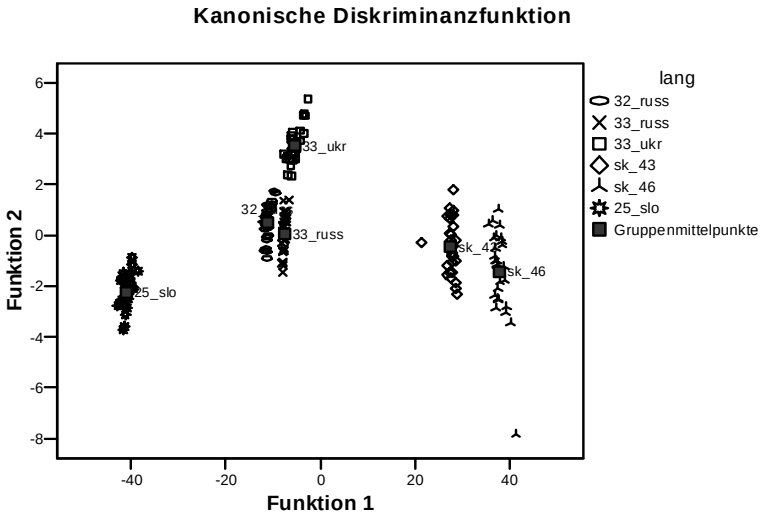


Abb. 11: Graphische Darstellungen der Diskriminanzfunktionen

Es liegt nahe, vor dem Hintergrund dieser Befunde eine Interpretation der Parameterwerte für *K* und *M* zu versuchen. Diese ergeben sich als

(9) $$K = \frac{M \cdot I}{m_1} \quad \text{bzw.} \quad K = \frac{M \cdot I}{m_1 - 1}$$

sowie

(10) $$M = \frac{m_1 \cdot K}{I} \quad \text{bzw.} \quad M = \frac{K \cdot (m_1 - 1)}{I}.$$

Sobald sich somit Anhaltspunkte für eine Interpretation von *K* oder *M* ergeben, wäre über die Bezugnahme auf den Mittelwert der Verteilung eine vollständige Parameterinterpretation möglich. Diesbezügliche in Arbeit befindliche Untersuchungen weisen auf eine besondere Rolle der ersten Häufigkeit (p_1) hin, doch muss dieses Thema einer eigenen, zukünftigen Untersuchung und Darstellung vorbehalten bleiben.

9. Zusammenfassung

Aufgrund der in der vorliegenden Untersuchung erhaltenen Ergebnisse lassen sich resümierend die folgenden wichtigsten Punkte zusammenfassen.

1. Die Graphemhäufigkeiten in allen bislang untersuchten slawischen Sprachen bzw. Stichprobenserien folgen dem Modell der negativen hypergeometrischen Verteilung (NHG);

2. Wie sich beim Versuch einer Interpretation der Parameter der NHG herausstellt, ist eine vermeintlich für eine Sprache repräsentative Korpusanalyse allein nicht ausreichend, um zu gesicherten Ergebnissen zu kommen:

   a. Erstens gilt ohnehin in der Sprache die Relation zwischen Stichprobe (Korpus) und Grundgesamtheit (Sprache) nicht, so dass es keine wenn auch noch so große Stichprobe geben kann, welche als „repräsentativ“ für eine Sprache angesehen werden könnte;
   b. Zweitens gibt es offensichtlich innerhalb einer Sprache sich nur in zahlreichen Stichproben manifestierende Mechanismen, die innerhalb einer Sprache synergetische Ausgleichmechanismen schaffen, und die auf spezifische Art und Weise die Häufigkeitsverhältnisse regeln.

3. Um vorliegende Trends zu beobachten und zu sichern, müssen in den einzelnen Stichproben Ausreißer kontrolliert und gegebenenfalls aus den einzelnen Analysen eliminiert werden, damit der Trend beobachtbar bleibt bzw. wird; das Vorkommen von Ausreißern dürfte im konkreten Fall auf zu kleine Stichproben zurückzuführen sein, die nicht die notwendige Sta-

bilität aufweisen – hier sind weiterführende Untersuchungen zur notwendigen Stichprobengröße notwendig und bereits in Arbeit.

4. Es gibt eine übersprachliche (sprachübergreifende) und einzelsprachliche Tendenzen der Regelung von Buchstabenhäufigkeiten:
   a. übersprachliche Tendenzen äußern sich für die bislang untersuchten slawischen Sprachen vor allem darin, dass sich die Abhängigkeiten der Parameter des bislang als am besten geeigneten Modells (NHG) auf ein gemeinsames Regressionsmodell zurückführen lassen, aus dem die einzelnen Sprachen als Spezialfälle hervorgehen.
   b. sprachspezifische Tendenzen äußern sich in den jeweils spezifischen Parameterwerten des bislang am besten geeigneten Modell (NHG) sowie der Möglichkeit, diese mit hinreichendem Erfolg als Diskriminanzfaktoren einsetzen zu können.
   c. Offenbar ist ungeachtet der einzelsprachlichen und übersprachlichen Tendenzen die Einordnung der einzelnen Sprachen in das Gesamtmodell von zusätzlichen „lokalen“ Randbedingungen überlagert, was sich im Falle des Ukrainischen zeigt; hier werden weitere Analysen darauf ausgerichtet sein müssen, diese Randbedingungen zu konkretisieren und gegebenenfalls bei der Modellbildung in Betracht zu ziehen. Absehbar ist, dass sich bei der Berücksichtigung weiterer – zumal auch nicht-slawischer – Sprachen Modifikationen des oben beschriebenen Modells ergeben; inwiefern sich das allgemeine Modell dabei aufrecht erhalten lässt, oder inwiefern es zur Option sprachtypologischer Differenzierungen kommen wird, können nur weitere systematische, weitaus umfassendere Untersuchungen zeigen.

5. Es muss kontrolliert werden, ob das Abweichen einzelner Sprachen wie etwa des Ukrainischen von der allgemeinen Tendenz nicht etwa (auch) durch die rein rechnerische Schätzung der Modellparameter bedingt ist; dazu wäre jedoch eine inhaltliche Interpretation der Parameter notwendig, die bislang nicht vorliegt; erste Hinweise deuten freilich darauf hin, dass sich die Parameterwerte (a) auf den Mittelwert, und (b) auf die erste Häufigkeit der jeweiligen Verteilungen beziehen lassen – diesbezügliche Untersuchungen laufen bereits.

Literatur

Best, K.-H. (2004/05): „Laut- und Phonemhäufigkeiten im Deutschen,“ in: *Göttinger Beiträge zur Sprachwissenschaft* (10/11), 21-32.

Best, K.-H. (2005): „Buchstabenhäufigkeiten im Deutschen und Englischen,“ in: *Науковий вісник Чернівеского університета*, вип. 231; 119-127.

Coulmas, F. (1981): *Über Schrift*. Frankfurt/M.
Derrida, J. (1974): *Grammatologie*. Frankfurt/M. (1. Auflage original frz. 1967, 9. Auflage 2004)
Dürscheid, Ch. (32006): *Einführung in die Schriftlinguistik*. 3., überarb. und ergänzte Aufl. Göttingen [= Studienbücher zur Linguistik, 8].
Glück, H. (1987): *Schrift und Schriftlichkeit. Eine sprach- und kulturwissenschaftliche Studie*. Stuttgart.
Grzybek, P. (2007a): „What a Difference an ‚E' Makes." [In print]
Grzybek, P. (2007b): „Zur Systematik der Untersuchung von Buchstabenhäufigkeiten im Deutschen". [In Druck]
Grzybek, P.; Kelih, E. (2003a): „Graphemhäufigkeiten (am Beispiel des Russischen) Teil I: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen", in: *Anzeiger für Slavische Philologie* (31), 131-162.
Grzybek, P.; Kelih, E. (2003b): „Grapheme Frequencies in Slovene." In: Benko, V. (ed.): *Slovko 2003*. Bratislava. [Ms.]
Grzybek, P.; Kelih, E. (2005a): „Graphemhäufigkeiten im Ukrainischen. Teil I: Ohne Apostroph." In: Altmann, G.; Levickij, V.; Perebejnis, V. (eds.): *Problemi kvantitativnoi lingvistiki – Problems of Quantitative Linguistics*. Černovici, 159-179.
Grzybek, P.; Kelih, E. (2005b): „Towards a General Model of Grapheme Frequencies in Slavic Languages." In: Garabík, R. (ed.): *Computer Treatment of Slavic and East European Languages*. Bratislava, 73-87.
Grzybek, P.; Kelih, E.; Altmann, G. (2004): „Häufigkeiten russischer Grapheme. Teil II: Modelle der Häufigkeitsverteilung", in: *Anzeiger für Slavische Philologie* (32), 25-54.
Grzybek, P.; Kelih, E.; Altmann, G. (2005a): „Graphemhäufigkeiten (am Beispiel des Russischen). Teil III: Die Bedeutung des Inventarumfangs – eine Nebenbemerkung zur Diskussion um das ‚ë'", in: *Anzeiger für Slavische Philologie* (33), 117-140.
Grzybek, P.; Kelih, E.; Altmann, G. (2005b): „Graphemhäufigkeiten im Slowakischen. (Teil I: Ohne Digraphen)." In: Nemcová, E. (Hrsg.): *Philologia actualis slovaca*. [In Druck]
Grzybek, P.; Kelih, E.; Altmann, G. (2006): „Graphemhäufigkeiten im Slowakischen. Teil II: Mit Digraphen." In: Kozmová, R. (Hrsg.): *Sprache und Sprachen im mitteleuropäischen Raum*. Trnava, 661-664.
Grzybek, P.; Mačutek, J.; Stadlober, E.; Wimmer, G. (2007): „Sample Size Estimation in Linguistics – A New Approach." [In print]
Hajnšek-Holz, M.; Jakopin, P. (1996): *Odzadnji slovar slovenskega jezika po Slovarju Slovenskega Knjižnega Jezika*. Ljubljana.

Harris, R. (2005): „Schrift und linguistische Theorie.“ In: Grube, G.; Kogge, W.; Krämer, S. (Hrsg.): *Schrift. Kulturtechnik zwischen Auge, Hand und Maschine*. (Kulturtechnik). München, 61-80.

Jakopin, P. (1995): „Nekaj števílk iz Slovarja Slovenskega Knjižnega Jezika“, in: *Slavistična revija* (43), 3, 341-375.

Jakopin, P. (2002): *Entropija v slovenskih leposlovnih besedilih*. Ljubljana.

Jakopin, P. (2003): Nekaj zanimivosti iz besedelnega korpusa Nova beseda, in: *Jezikoslovni zapiski* (9/2), 145-152.

Kleinbaum, D.G.; Kupper, Lawrence L.; Muller, K.E. (^{3}1998): Applied regression analysis and other multivariable methods. Pacific Grove, 3rd ed., rev.

Köhler, R. (2005): „Synergetic linguistics“. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (Hrsg.): *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook*. Berlin u.a., 760-774. [= Handbücher zur Sprach- und Kommunikationswissenschaft, 27]

Poníž, D. (1974): *Slovenski jezik – literatura – računalniki*. Maribor.

peter.grzybek@uni-graz.at
emmerich.kelih@uni-graz.at
e.stadlober@tugraz.at