Anzeiger für Slavische Philologie
(XXXI) 2003, 131-162

Graphemhäufigkeiten (am Beispiel des Russischen)

Teil I: Methodologische Vor-Bemerkungen und Anmerkungen zur Geschichte der Erforschung von Graphemhäufigkeiten im Russischen[1]

Peter Grzybek, Emmerich Kelih (Graz)

0. Einleitung: Methodologische Vorbemerkungen

Der vorliegende Aufsatz stellt den Auftakt einer Reihe von Untersuchungen dar, in denen am Beispiel der Analyse russischer Graphemhäufigkeiten die prinzipielle Vorgangsweise der Quantitativen Sprach- und Textanalyse veranschaulicht werden soll.[2]

Anlass zu einer solchen grundsätzlichen und systematischen Darstellung gibt nicht zuletzt die Tatsache, dass sich im Zusammenhang mit den Entwicklungen der Computertechnologie in den letzten zwei Jahrzehnten im Bereich der Sprach- und Textwissenschaft zwei Richtungen etabliert haben, die üblicherweise unter der Bezeichnung ‚Computerlinguistik' (vgl. z.B. Bátori et al. 1989) und ‚Korpuslinguistik' (vgl. Lenz 2000) firmieren. Bei aller Unterschiedlichkeit der Interessen und Herangehensweisen der einzelnen (Teil-) Disziplinen lassen sich als Konsequenzen der technischen Entwicklungen zwei wesentliche Feststellungen herausfiltern:

---

[1] Die vorliegende Untersuchung entstand im Zusammenhang mit dem FWF-Projekt P-15485 »Wortlängen(häufigkeiten) in Texten slawischer Sprachen« (vgl. <http://www-gewi.uni-graz.at/quanta>).

[2] Unter Vermeidung der notwendigen, jedoch an anderer Stelle systematisch zu führenden Diskussion zur differenzierten Verwendung der Termini ‚Buchstabe', ‚Graph', ‚Graphem', u.a. – unter Einschluss solcher (je nach Ausgangsdefinition unterschiedlich verwendeter) Bezeichnungen wie ‚Digraph', ‚Diagraphem' u.a.m. – sollen im vorliegenden Text unter „Graphemen" die den darzustellenden Untersuchungen undifferenziert zugrunde gelegten Einheiten des (jeweiligen) Alphabetbestands verstanden werden. – Ebenso ausgeblendet bleibt eine systematische Darstellung des russischen Graphembestandes – zumal in dessen historischer Entwicklung (vgl. hierzu u.a. Spraul 1999, u.a.m.).

a) zum einen hat sich aufgrund der wesentlich erleichterten technischen Handhabbarkeit die relativ einfache Möglichkeit der Bearbeitung größerer Datenmengen, vor allem auch ganzer Korpora, durch auch nur ansatzweise technisch versierte NutzerInnen ergeben;
b) zum anderen ist aufgrund dieser Entwicklungen – und insbesondere auch im Zusammenhang mit den unter (a) genannten Faktoren – die Grenze zwischen qualitativer und quantitativer Sprach- bzw. Textanalyse zunehmend verschwommen.

Vor diesem Hintergrund liegt es nahe, eingangs im Hinblick auf die oben angesprochene Fragestellung die eigentlichen Forschungsinteressen und damit einhergehenden Vorgehensweisen der Quantitativen Sprach- und Textanalyse klar zu konturieren.

Geht man operational von der Gegenüberstellung einer qualitativen und einer quantitativen Sprach- und Textwissenschaft aus, so ergibt sich als vorrangige Aufgabe eines quantitativen Zugangs die Auseinandersetzung mit der statistischen Beschaffenheit sprachlicher bzw. textueller Gegebenheiten. Diese Annahme impliziert, dass man qualitative Eigenschaften und Strukturen von Sprache(n) und Text(en) messbar macht und im Anschluss daran mit anderen Mitteln versucht, dem (scheinbar) irregulären und vagen Charakter sprachlicher Eigenschaften gerecht zu werden. In diesem Zusammenhang wird üblicherweise von folgenden Postulaten ausgegangen (vgl. Altmann 1972: 2f):

a.) Sprache(n) und Text(e) werden nicht (nur) als Ansammlung individueller Wesensmerkmale, sondern zugleich auch als eine Massenerscheinung angesehen, wobei es gilt, die in ihnen vorliegenden Tendenzen (Regularitäten, Gesetzmäßigkeiten) nachzuweisen;
b.) Die Quantifizierung von Sprach- und Texteigenschaften ist ein Verfahren, mit dessen Hilfe man die Eigenschaft des untersuchten Objektes so operationalisiert, dass man ihre Ausprägungen in einer Zahlenmenge abbildet und somit auch Beziehungen zwischen den Objekten durch numerische Relationen ausdrücken kann.[3]

[3] Aus dem bisher Gesagten und vor allem unter Berücksichtigung und Betonung des dynamischen Charakters des Objekts ‚Sprache' und ‚Text' sowie der Anwendung wahrscheinlichkeitstheoretischer (stochastischer) Methoden geht die Notwendigkeit hervor, die quantitative Sprach- und Textanalyse von der (auf algebraischen und mengentheoretischen Konzeptionen basierenden) mathematischen Linguistik zu unterscheiden, die von deterministischen Modellen ausgeht und versucht, sprachliche Phänomene auf der Basis von Algebra und Algorithmentheorie zu beschreiben (vgl. dazu auch Gladkij/Mel'čuk 1973: 21ff) – einen Anspruch auf die Erklärung sprachlicher Phänomene erhebt die mathematische Linguistik dabei in der Regel nicht.

In diesem Sinne dient die *quantitative Linguistik* als *Hilfsdisziplin* der *qualitativen Linguistik*: „Die Zahlen sind nicht das Endziel der Forschung, sondern dienen nur als Indikatoren bestimmter Qualitäten. Die Sprachwissenschaft ist nicht an reinen Zahlen interessiert, sondern an der Dechiffrierung des Mechanismus der Sprache, dessen Bestandteile sich aber häufig am besten durch numerische Relationen ausdrücken lassen“ (Altmann 1972: 7).

Natürlich ist die Anwendung statistischer Methoden nicht allein dem Kernbereich der quantitativen Sprach- und Textanalyse vorbehalten; vielmehr findet sie durchaus auch in anderen, etablierten Bereichen der Linguistik und Textwissenschaft, wie etwa der Korpuslinguistik und der Computerlinguistik, Anwendung. Dabei kommt der Quantifizierung im Selbstverständnis der Quantitativen Sprach- und Textanalyse allerdings ein wesentlicher Beitrag zur Theoriebildung zu, was sie von der Korpus- und Computerlinguistik trennt:

- So versteht sich die *Korpuslinguistik* – die sich ja de facto über ihren Gegenstandsbereich definiert – als eine zwar theoretisch untermauerte, dennoch aber dezidiert daten-orientierte, empirische Form der Wissenschaft: Zwar ist es ihr Anliegen, auf der Grundlage hinreichend großer Textmengen (Korpora) begründete Theorieentwürfe zu machen, doch hat sie keinen direkten Einfluss auf die eigentliche Theoriebildung – dazu ist sie vielmehr auf die eine oder andere Art der Quantifizierung angewiesen.
- Der *Computerlinguistik* hingegen – die sich ja nur auf der untersten Stufe über die der Analyse zugrunde gelegten technischen Voraussetzungen definiert – geht es in letzter Konsequenz um eine möglichst exakte Deskription und/oder die darauf aufbauende Simulation sprachlicher Prozesse. Zwar versteht sie mitunter genau dies als ihren eigentlichen Theoriebeitrag, doch kommt auch sie hierbei nicht ohne Bezugnahme auf statistische Verfahren aus.

Insofern lässt sich der Stellenwert der Quantitativen Linguistik im Grunde genommen wie folgt darstellen:

Computerlinguistik	⇔	Quantitative Linguistik	⇔	Qualitative Linguistik

Dabei zeichnet sich die Untersuchung sprachlicher Phänomene auf der Grundlage von quantitativen Methoden durch eine klar definierte Vorgangsweise aus. Nach Altmann (1972, 1973) sollte eine quantitativ-linguistische Untersuchung nach folgenden Schritten erfolgen:

I. Der erste Schritt beinhaltet die **Aufstellung linguistischer Hypothesen** in Bezug auf die zu verfolgende Fragestellung; wichtig ist hierbei vor allem, dass die Hypothese von empirischer Relevanz und in der Folge empirisch überprüfbar ist. In Anbetracht der prinzipiellen Unmöglichkeit, für einen infiniten Objektbereich endgültig verifizierbare Hypothesen postulieren zu können, geht es um die Aufstellung intersubjektiv nachvollziehbarer Wahrscheinlichkeitshypothesen.
II. Der zweite Schritt besteht in der **Übersetzung** der zuvor formulierten Hypothese in die Sprache der Statistik, d.h. die statistische Formulierung der Hypothese.
III. Die **empirische Überprüfung** als dritter Schritt impliziert einerseits die Gewinnung bzw. Zusammenstellung einer angemessenen Stichprobe, welche einigermaßen repräsentative Aussagen über die zu beantwortende Fragestellung zulässt, andererseits die Wahl angemessener statistischer Methoden und Testverfahren. Dieser Schritt zielt somit im wesentlichen auf eine Beantwortung der Frage, ob die in eine statistische Hypothese transformierte linguistische Hypothese beibehalten werden kann oder verworfen werden muss.
IV. Die **Entscheidung** über Annahme oder Ablehnung der aufgestellten Hypothese ist Gegenstand des vierten Schritts, der eine **statistische Interpretation der Ergebnisse** im Hinblick auf die anfangs aufgestellte Hypothese vorsieht.
V. Der fünfte und letzte Schritt schließlich beinhaltet die Rück-Übersetzung der erhaltenen statistischen Ergebnisse, d.h. die **linguistische Interpretation** des Resultats der Entscheidung, die es letztlich in eine allgemeine Sprachtheorie zu integrieren gilt.

Zu dieser idealen Abfolge ist zu sagen, dass zahlreiche sprach- und textwissenschaftliche Arbeiten, die sich selbst als „quantitativ" bezeichnen, nicht diesem in sich schlüssigen Abfolgeschema entsprechen. So findet man auch heutzutage noch nicht wenige Arbeiten, die sich ausschließlich auf der zweiten Ebene bewegen: dabei werden entweder Häufigkeiten bzw. Proportionen von ausgewählten Elementen präsentiert, ohne dass der Zusammenhang zu einer irgendwie gearteten (zuvor aufgestellten) Hypothese erkennbar wäre; oder aber es werden Resultate quasi als Endergebnis präsentiert, die jedoch aufgrund der fehlenden statistischen Verfahren auf einer intuitiven Ebene stehen bleiben oder gar überhaupt keiner weiteren Interpretation unterzogen werden – zumal dann, wenn zuvor keine entsprechenden Hypothesen aufgestellt wurden (vgl. Altmann 1973: 218ff). Solche Arbeiten sind allein insofern nicht vollkommen wertlos, als sie zumindest nützliches Material für weitere Untersuchungen bereitstellen.

Vor dem Hintergrund der obigen methodologischen Bemerkungen soll die vorliegende Teilstudie als Exemplifizierung des Gesagten eine systematische Untersuchung der Vorkommenshäufigkeit russischer Grapheme zum Gegenstand haben. Dazu ist einschränkend zu sagen, dass es ausschließlich um das Vorkommen einzelner Grapheme gehen soll und nicht etwa um Graphemkombinationen oder um das Vorkommen von Graphemen in bestimmten Positionen.[4] Ohne im hier gegebenen Kontext eine hierarchische Nach- oder Nebenordnung zwischen lautlicher und schriftlicher Repräsentationsform zu postulieren, und ohne Rücksichtnahme auf die Adäquatheit der möglichen wechselseitigen Abbildbarkeit beider Formen, dienen als Gegenstand der Untersuchung ausschließlich die schriftlichen Repräsentationsformen. Auch wenn diese graphematische Ebene der Sprache häufig im Vergleich etwa zur morphologischen, lexikalischen, syntaktischen Ebene als eher „niedrig" angesehen wird, handelt es sich bei der Analyse des Graphembestands einer Sprache und der Vorkommenshäufigkeit der diese Ebene konstituierenden Elemente ohne Zweifel um eine wichtige Frage im Hinblick auf deren systemisch-synergetisches Funktionieren.

In der hier vorgelegten Untersuchung wird die graphematische Ebene der russischen Sprache jedoch nicht nur als an und für sich relevantes Forschungsobjekt betrachtet; vielmehr geht es auch und gerade darum, an diesem ausgewählten Gegenstandsbereich methodologische Grundprinzipien der quantitativen Sprach- und Textwissenschaft zu veranschaulichen, um so einerseits den Status vorliegender Untersuchungen besser einschätzen zu können, und um andererseits einen Weg aufzuzeigen, der dem Anspruch einer umfassenden Theoriebildung Genüge zu leisten vermag.

Die Behandlung der genannten Fragestellung kann nicht in einem einzigen Teil geleistet werden. Während der hier vorgelegte erste Teil primär eine historisch ausgerichtete Darstellung bislang vorgelegter Untersuchungen zu dieser Frage darstellt[5], wird es in einem zweiten Teil darum gehen, theoretische Modelle für die Vorkommenshäufigkeit russischer Grapheme zu ana-

---

[4] Unter Konzentration auf die Untersuchung der Vorkommenshäufigkeit der einzelnen Einheiten des gesamten Graphembestandes bleiben also auch zahlreiche Untersuchungen ausgeklammert, die etwa den prozentualen Anteil von Konsonanten und Vokalen gegenüberstellen, oder auch Untersuchungen wie jene von Markov (1913), der darüber hinausgehend am Beispiel von Puškins *Evgenij Onegin* auf wahrscheinlichkeitstheoretischer Basis die Übergangswahrscheinlichkeiten zwischen Konsonanten und Vokalen berechnete.

[5] Die vorliegende Darstellung ist um Vollständigkeit der Darstellung des bislang Geleisteten bemüht; eine Gewährleistung, dass in der Tat alle jemals erbrachten Frequenzuntersuchungen zum Russischen Eingang in diese Darstellung gefunden haben, kann dies natürlich nicht sein. Die Autoren wären insofern für Hinweise auf weitere Untersuchungen und Lücken in der Darstellung dankbar.

lysieren, die bislang in der einschlägigen Diskussion zum Tragen gekommen sind (vgl. Grzybek/Kelih/Altmann 2004). Diese theoretischen Modelle werden in der weiteren Folge zu den Ergebnissen eigener empirischer Untersuchungen in Beziehung zu setzen, an anderen Sprachen zu überprüfen, dabei gegebenenfalls zu erweitern und auf jeden Fall zu interpretieren sein.

2. Zur Geschichte der Untersuchung von Graphemhäufigkeiten im Russischen

Aufgrund der schlechten Zugänglichkeit und schweren Verfügbarkeit der meisten der bislang erfolgten einschlägigen Untersuchungen sollen in einem ersten Schritt die entsprechenden statistischen Befunde schlicht und einfach im Kontext ihres Entstehens dargestellt werden. Dabei wird schnell zu sehen sein, dass eine entsprechende „einfache" Darstellung recht bald zu einer Diskussion der mit diesen Erhebungen verbundenen statistischen Analysen führt, die in den meisten Fällen entweder auf die Frage der Validität (Repräsentativität) der erhaltenen Daten oder auf die Frage der Vergleichbarkeit verschiedener Untersuchungen zielt. Insofern leitet die folgende historische Aufarbeitung relativ logisch zu der im zweiten Teil im Vordergrund stehenden Frage nach der theoretischen Modellierung der Vorkommenshäufigkeit russischer Grapheme über.

2.1. Budilovič (1883)

In historischen Darstellungen zur Geschichte der Quantitativen Linguistik in Russland wird gern auf die Daten von Anton Budilovič als der frühesten russischen Graphemstatistik verwiesen. In der Tat sind in seiner kirchenslawischen Grammatik (1883: 67, 97) prozentuale Vorkommenshäufigkeiten – getrennt nach Vokalen und Konsonanten – angeführt. Allerdings handelt es sich nicht, wie in verschiedenen Darstellungen angenommen wird, um von Budilovič selbst erhobene Daten, sondern sie stammen von August Schleicher, der diese erstmals gut drei Jahrzehnte vorher in seiner *Formenlehre der kirchenslawischen Sprache* (1852: 20f) präsentiert hatte; dies wollte Schleicher als Ergänzung zu den von Förstemann (1846, 1852) zuvor erhobenen Daten zum Deutschen, Griechischen, Lateinischen und Gotischen verstanden wissen. Tab. 1 enthält die entsprechenden Daten, die in zwei Fällen (**ц**, **ѵ**) Widersprüche zwischen den (zuerst genannten) Daten von Schleicher und Budilovič aufweisen.

Tab. 1: Graphemhäufigkeiten im Kirchenslawischen nach Schleicher / Budilovič

и	20,4	т	10,7	ш	5,1
ъ	13,9	в	9,1	г	4,3
є	13,4	н	8,7	к	3,9
о	13,4	с	8,7	ж	3,9
а	12,7	ж	6,9	ь	3,5
ѣ	6,1	р	6,8	п	3,2
ь	6	д	6,4	ц	0,8 / 2,2
ѧ	4,3	м	6,3	з	1,8
ѫ	4,3	л	6,2	х	1,5
ъı	3			ѵ	2,2 / 0,8
оѵ	2,5				

2.2. Ol'chin (1907)

In Anbetracht der Tatsache, dass die Daten von Schleicher bzw. Budilovič auf Untersuchungen zum Kirchenslawischen beruhen, dürfte die Erhebung von P. Ol'chin (1907) die erste statistische Untersuchung russischer Graphemhäufigkeiten darstellen. Diese Untersuchung wurde primär aus praktischen Motiven durchgeführt: sie stellt nämlich einen Versuch dar, Vorkommenshäufigkeiten russischer Grapheme zum Zwecke der Optimierung der stenographischen Notation zu bestimmen. Als Textgrundlage der Erhebung dienten dem Autor sechs verschiedene Stichproben:

1. 621 Wörter aus einer politischen Rede
2. 735 Wörter aus einer politischen Rede
3. 467 Wörter aus einer politischen Rede
4. 900 Wörter aus einem Buch zur Kindererziehung
5. 820 Wörter aus einem Lehrbuch zur russischen Sprache
6. 894 Wörter aus einem Buch von S. Krasevič (*Zemlja i nebo*)

Untersucht wurde die Vorkommenshäufigkeit von 29 Graphemen, ohne Berücksichtigung von Groß- oder Kleinschreibung:
АБВГДЕЖЗИКЛМНОПРСТУФХЦЧШЩЙ.

Für diese Grapheme wurde die Vorkommenshäufigkeit sowohl in den einzelnen Texten (I-VI) als auch insgesamt angegeben.[6] Eine Reihe von Ungereimtheiten und Unzulänglichkeiten lässt jedoch die Verwertbarkeit der Ergebnisse zumindest für andere als die vom Autor angestrebten Ziele als zweifelhaft erscheinen:

1. Es werden nicht alle Grapheme des (damaligen) russischen Alphabets analysiert, weil der Autor die seiner Ansicht nach für stenographische Zwecke nicht notwendigen Einheiten auslässt; so bleiben zum einen die Halbvokale Ъ und Ь unberücksichtigt, zum anderen die in der alten Orthographie verwendeten Grapheme Ѣ, Ѳ, Ѵ, Э, und І;
2. der Autor führt in seinen Erläuterungen nur 26 der oben aufgeführten Grapheme an, in den Tabellen sind jedoch (unter zusätzlicher Berücksichtigung von Я, Ю, Ы) die Werte für 29 Grapheme wiedergegeben;
3. aus der Analyse wurden sämtliche Präfixe sowie ihnen entsprechende Präpositionen ausgeschlossen;
4. die für die einzelnen Grapheme in den Tabellen angegebenen Gesamtsummen entsprechen in mehreren Fällen nicht der Summe der angegebenen Häufigkeiten in den Einzeltexten, wobei sich natürlich nicht rekonstruieren lässt, welche der angegebenen Werte jeweils falsch sind; da dem Autor aber zudem auch bei einfachen Additionen (z.B. der summierten Vorkommenshäufigkeiten der Vokale) Fehler unterlaufen, ist es wahrscheinlicher, dass die Angaben für die einzelnen Texte stimmen, nicht aber die Ergebnisse der Additionen.

Tab. 2 enthält sowohl die vom Autor angegebenen als auch die für die vorliegende Darstellung nachberechneten Summen und Teilsummen.

[6] Ebenfalls untersucht wurden die im hier gegebenen Kontext nicht zu diskutierenden Vorkommenshäufigkeiten konsonantischer Zweier-, Dreier-, und Viererkombinationen, sowie die Vorkommenshäufigkeiten von aus mehr als einem Graphem bestehenden Präfixen und Präpositionen.

Tab. 2: Graphemhäufigkeiten nach Ol'chin (1907)

	I	**II**	**III**	**IV**	**V**	**VI**	**Σ** (Ol'chin)	**Σ** (korr.)
Е	278	394	188	575	483	457	2475	*2375
О	341	411	225	535	522	426	2460	2460
И	201	300	183	320	440	226	1710	*1670
А	258	294	200	316	305	287	1660	1660
Н	275	316	118	320	376	251	1656	1656
Т	181	283	146	343	267	282	1502	1502
С	179	238	129	189	238	258	1231	1231
В	157	175	119	165	228	237	1081	1081
Р	151	173	103	174	217	130	948	948
Л	135	150	94	102	194	195	880	*870
К	100	147	94	221	151	138	851	851
М	123	136	80	146	167	165	817	817
Д	134	108	61	175	140	120	738	738
У	134	100	50	126	115	121	646	646
Я	89	104	38	93	99	93	516	516
Ы	44	81	74	50	84	86	419	419
Г	55	75	46	65	91	84	416	416
Б	56	40	40	105	73	43	357	357
П	41	62	33	57	84	60	337	337
Ч	26	56	24	74	68	69	317	317
Й	34	34	21	43	66	31	229	229
Х	41	30	37	34	45	32	219	219
Ж	8	31	15	56	41	50	201	201
З	12	35	31	30	54	36	198	198
Ю	21	24	10	47	49	27	178	178
Ш	13	22	15	35	35	55	174	*175
Щ	20	13	9	25	22	11	100	100
Ц	5	18	8	7	43	6	87	87
Ф	4	17	3	16	10	0	50	50

2.3. Morozov

Eine erste alle Buchstaben des Russischen (der damaligen Zeit) erfassende Statistik stammte von Nikolaj A. Morozov (1915), einem ehemaligen dem Terrorismus nahe stehenden Volkstümler (narodnik), der später (ab 1932) Ehrenmitglied der Akademie der Wissenschaften der UdSSR werden sollte. In seiner ein Jahr später (1916) in Petersburg auch als eigenständige Broschüre herausgegebenen Arbeit war er seinen eigenen Aussagen zufolge primär an der Bestimmung individual-stilistischer Merkmale interessiert. Dieses – seinen Angaben zufolge schon Mitte der 80er Jahre des 19. Jahrhunderts vorliegende – Interesse zielte allerdings im Grunde genommen eher auf „allgemeine stilometrische Gesetze" (ebd., 97). Wenn, so seine Überlegung, in der Natur und im gewöhnlichen Leben die vielfältigsten, scheinbar zufälligen, Erscheinungen in einem beträchtlichen Maße „gesetzmäßigen Charakter" haben – warum soll das nicht auch auf den Bereich der Sprache zutreffen?

Als eine erste Veranschaulichung seiner Überzeugung, dass „in unseren menschlichen Sprachen all deren Elemente eine bestimmte Proportion haben" (ebd. 97), und dass „statistische Gesetzmäßigkeiten nicht nur in sich häufig wiederholenden Erscheinungen der Natur und des öffentlichen Lebens, sondern auch in den Erscheinungen unserer Umgangs- und Schriftsprache existieren" (ebd., 110), führte Morozov eine Häufigkeitsliste russischer Buchstaben an, ohne allerdings zu sagen, woher die entsprechenden Zahlen stammen bzw. worauf sie basieren.

Tab. 3: Graphemhäufigkeiten nach Morozov (1915)

а	540	и	470	р	375	ы	160
б	160	і	160	с	420	ь	160
в	335	к	250	х	85	ѣ	160
г	160	л	250	ц	85	э	45
д	330	м	250	ч	125	ю	85
е	550	н	250	ш	80	я	210
ж	85	о	665	щ	75	ѵ	40
з	125	п	200	ъ	375	й	125

Einen Vergleich mit anderen Stichproben führte Morozov allerdings nicht durch, um seine Annahme der Gesetzmäßigkeit der Vorkommenshäufigkeit von Buchstaben zu belegen, so dass diese Annahme aufgrund seiner Untersuchung allein nicht weiter verfolgt werden kann.

2.3. Proskurnin (1933)

Konkret-praktischen Zielen diente hingegen eine 1933 von N. Proskurnin vorgelegte Studie, die das Ergebnis von Untersuchungen war, die im Kontext mit den neu geschaffenen bzw. modifizierten Alphabeten in der UdSSR der 20er und 30er Jahre stand. Hier ging es um die optimierte Herstellung von Drucktypen, die sich an der Vorkommenshäufigkeit der zu druckenden Buchstaben orientieren sollte. Im Gegensatz zu vor-revolutionären Zeiten war dies in der frühen UdSSR zu einer staatlichen Angelegenheit geworden, die einerseits mit Fragen der Standardisierung, andererseits aber mit ökonomischen Gesichtspunkten zusammenhing

Proskurnin (1933: 75) veröffentlichte zu diesem Zweck die Ergebnisse einer Auszählung von mehr als 1 Million gedruckten Zeichen, differenziert nach Klein- und Großbuchstaben, sowie unter Einbeziehung von Interpunktionszeichen und Ziffern. Tab. 4 gibt die absoluten Werte der 33 russischen Buchstaben wieder; ebenso angeführt sind die entsprechenden prozentualen Häufigkeiten, die insofern von den bei Proskurnin (1933: 75) angeführten abweichen, als sie sich ausschließlich auf die Summe der genannten 33 Buchstaben des russischen Alphabets beziehen. In Ergänzung zu den nach Groß- und Kleinschreibung differenzierten Häufigkeiten sind in der Tab. 4 auch die zusammengefassten Frequenzen enthalten.

2.4. Prachov (1946)

In den 40er Jahren führte P.V. Prachov (1946) im Zusammenhang mit der Optimierung drei- und vierreihiger Telegraphen-Maschinen eine später kaum noch erwähnte Untersuchung zur Vorkommenshäufigkeit russischer Grapheme durch. Bekannt geworden sind die Daten allerdings, weil in unterschiedlichen Zusammenhängen immer wieder auf sie verwiesen wurde, ohne dass die ursprüngliche Quelle dabei erwähnt wurde. Zu finden sind sie z.B. bei Charkevič (1955: 233f), der die Daten im Zusammenhang mit einer Optimierung des Morse-Codes für die russische Sprache diskutiert[7]; ohne Angabe der Quelle finden die Daten von Prachov auch Erwähnung z.B. bei Jaglom/ Jaglom (1960), die sie anderen Autoren, nämlich Lebedev und Garmaš zuschreiben, auf die unten noch einzugehen sein wird.

[7] Bei der detektivischen Suche nach dem Ursprung dieser Daten war uns V.V. Kromer (Novosibirsk) sehr behilflich; ihm sei an dieser Stelle herzlich für seine Hilfsbereitschaft gedankt.

Tab. 4: Graphemhäufigkeiten nach Proskurnin (1933)

а	74432	7,57	**А**	628	4,01	75060	7,51
б	16575	1,69	**Б**	454	2,90	17029	1,70
в	43711	4,44	**В**	1565	10,00	45276	4,53
г	15518	1,58	**Г**	390	2,49	15908	1,59
д	29603	3,01	**Д**	594	3,80	30197	3,02
е	86713	8,82	**Е**	448	2,86	87161	8,72
ж	9646	0,98	**Ж**	91	0,58	9737	0,97
з	17184	1,75	**З**	362	2,31	17546	1,76
и	73675	7,49	**И**	725	4,63	74400	7,45
к	32624	3,32	**К**	1038	6,64	33662	3,37
л	41732	4,24	**Л**	290	1,85	42022	4,21
м	30490	3,10	**М**	702	4,49	31192	3,12
н	63147	6,42	**Н**	1363	8,71	64510	6,46
о	109011	11,08	**О**	1009	6,45	110020	11,01
п	26539	2,70	**П**	1480	9,46	28019	2,80
р	47022	4,78	**Р**	636	4,07	47658	4,77
с	53578	5,45	**С**	1364	8,72	54942	5,50
т	64102	6,52	**Т**	830	5,31	64932	6,50
у	24568	2,50	**У**	240	1,53	24808	2,48
ф	1742	0,18	**Ф**	142	0,91	1884	0,19
х	10545	1,07	**Х**	170	1,09	10715	1,07
ц	4388	0,45	**Ц**	93	0,59	4481	0,45
ч	14640	1,49	**Ч**	286	1,83	14926	1,49
ш	6743	0,69	**Ш**	88	0,56	6831	0,68
щ	4442	0,45	**Щ**	7	0,04	4449	0,45
ъ	331	0,03	**Ъ**	0	0,00	331	0,03
ы	19699	2,00	**Ы**	3	0,02	19702	1,97
ь	17269	1,76	**Ь**	12	0,08	17281	1,73
э	2777	0,28	**Э**	388	2,48	3165	0,32
ю	7239	0,74	**Ю**	32	0,20	7271	0,73
я	21668	2,20	**Я**	206	1,32	21874	2,19
й	11824	1,20	**Й**	7	0,04	11831	1,18
ё	382	0,04				382	0,04
	983559			15643		999202	

Charkevič ging bei der Verwendung der Daten von Prachov davon aus, dass zum Zwecke einer statistisch optimierten Codierung die Buchstaben in abnehmender Vorkommenshäufigkeit zu rangieren und mit den in aufsteigender

Rangreihenfolge zu sortierenden (als länger werdenden) Code-Bezeichnungen zu korrelieren sind. Die von ihm reproduzierten Daten von Prachov finden sich in der Tab. 5.

Tab. 5: Graphemhäufigkeiten nach Prachov (1946)

Graphem	f (%)	Graphem	f (%)	Graphem	f (%)	Graphem	f (%)
О	11,0	**В**	4,6	**Ы**	1,9	**Ж**	0,9
Е	8,7	**Л**	4,2	**З**	1,8	**Ю**	0,7
А	7,5	**К**	3,4	**Ь, Ъ**	1,7	**Ш**	0,7
И	7,5	**М**	3,1	**Б**	1,7	**Ц**	0,5
Т	6,5	**Д**	3,0	**Г**	1,6	**Щ**	0,4
Н	6,5	**П**	2,8	**Ч**	1,5	**Э**	0,3
С	5,5	**У**	2,5	**Й**	1,2	**Ф**	0,2
Р	4,8	**Я**	2,2	**Х**	1,1		

2.5. Lebedev/Garmaš (1958)

Mit dem Ziel der Optimierung von Nachrichtenübertragungen haben D.S. Lebedev und V.A. Garmaš 1958 die Vorkommenshäufigkeit von Graphemen und Graphemkombinationen untersucht. Ziel dieser Untersuchung war die Beschleunigung der Informationsübertragung. Der expliziten Aussage von Lebedev/Garmaš (1958: 68) zufolge, die selbst in diesem Zusammenhang an der Vorkommenshäufigkeit von Graphemkombinationen interessiert waren, galt seinerzeit die Einbuchstaben-Statistik für die russische Sprache als bekannt – auf welche Daten sich die Autoren mit dieser Aussage beziehen, wird von ihnen allerdings nicht gesagt. Spätere Autoren wie z.B. Jaglom/ Jaglom (1960) „zitieren“ zwar Einbuchstaben-Frequenzen unter direktem Verweis auf die Arbeiten von Lebedev/Garmaš (1958, 1959) – in diesen beiden Texten aber sind die entsprechenden Angaben nicht enthalten![8] Die Angaben von Jaglom/Jaglom (1960) decken sich allerdings nicht nur mit denen, die sich dann auch bei Andreeva et al. (1965) finden – sie stimmen vor allem, wie unten noch gezeigt werden wird, zu nahezu 100% mit denen von Prachov (1946) überein. Da weiter unten noch eine Darstellung dieser Angaben in

[8] Wie Uspenskij (2002/II: 942) schreibt, wurden die Unterlagen (d.h. die damals verwendeten Lochkarten und die Berechnungen) unmittelbar nach der Untersuchung vernichtet, ob auf der Basis der Mehrbuchstaben-Frequenzen jemals die Einbuchstaben-Frequenzen rekonstruiert wurden, wird nirgends explizit gesagt.

direktem Vergleich mit anderen Analysen erfolgen wird, kann an dieser Stelle auf eine weitere Diskussion dieser Daten verzichtet werden.

2.6. Belonogov/Frolov (1963)

Belonogov/Frolov (1963) haben mit dem Ziel der „Lösung theoretischer und praktischer Aufgaben im Zusammenhang mit der automatischen Informationsverarbeitung“ die Vorkommenshäufigkeit der 32 russischen Grapheme (ohne gesonderte Berücksichtigung des „ë“) auf der Basis eines Häufigkeitswörterbuchs berechnet, das zuvor auf der Basis von Fachtexten erstellt worden war. In die Auswertung gingen ca. 30.000 Wörter ein, deren Graphemsumme sich insgesamt auf ca. 200.000 belief. Angeführt wurden allerdings nur die relativen, nicht die absoluten Häufigkeiten. Tab. 6 gibt die von den Autoren angegebenen Werte in absteigender Rangreihenfolge wieder:

Tab. 6: Graphemhäufigkeiten nach Belonogov/Frolov (1963)

Rang	Graphem	p	Rang	Graphem	p
1	О	0,1047	17	Ч	0,0180
2	Е	0,0836	18	Ы	0,0179
3	А	0,0808	19	З	0,0174
4	Н	0,0723	20	Б	0,0170
5	И	0,0700	21	Ь	0,0168
6	Т	0,0625	22	Й	0,0158
7	Р	0,0584	23	Х	0,0132
8	В	0,0569	24	Г	0,0111
9	С	0,0466	25	Ж	0,0096
10	Д	0,0388	26	Ю	0,0063
11	П	0,0371	27	Ш	0,0050
12	М	0,0337	28	Щ	0,0035
13	К	0,0264	29	Ц	0,0029
14	Я	0,0249	31	Э	0,0017
15	Л	0,0248	30	Ф	0,0017
16	У	0,0202	32	Ъ	0,0003

2.7. Andreeva et al. (1965)

Andreeva et al. (1965) haben quantitative Untersuchungen zur Graphemhäufigkeit an russischen Texten aus dem Gebiet der Radioelektronik durchgeführt und zu diesem Zweck ein Korpus von insgesamt ca. 1 Million Wortformen analysiert. In diesem Zusammenhang haben sie auch erstmals vergleichend die Ergebnisse verschiedener Untersuchungen zur Vorkommenshäufigkeit von Graphemen in unterschiedlichen Textsorten tabellarisch zusammengefasst. Die folgende Aufstellung beinhaltet eine Zusammenschau der betreffenden Untersuchungen:

Abk.	**Autor(en)**	**Jahr**	**Textsorte**
GML	Gruppe »Mathematische Linguistik«	1961	Radioelektronik
A	Andreeva	1959	allg. wissenschaftlich-technisch
LG	Lebedev/Garmaš	1958	allgemein
BF	Belonogov/Frolov	1961	Fachtexte

In der Tab. 7 sind die von Andreeva et al. (1965) angeführten Ergebnisse hinsichtlich der 32 Grapheme des Russischen wiedergegeben; die aufsummierten relativen Häufigkeiten ergeben in allen Fällen nicht die Summe 1, weil in den einzelnen Untersuchungen Interpunktionszeichen, Gedanken- und Trennstriche, Klammern, etc. unterschiedlich berücksichtigt wurden.

Wie zu sehen ist, decken sich die Angaben, die sich auf die Untersuchungen von Lebedev/Garmaš beziehen, fast vollständig mit den oben angeführten Daten von Prachov (1946): Berücksichtigt man, dass die Summe der Prozentwerte zu den vermeintlichen Daten von Lebedev/Garmaš 82.60% beträgt, und dass Andreeva et al. (1965: 51) die Vorkommenshäufigkeit des Leertasten-Zwischenraums mit 17.5% beziffern, führt eine Umrechnung der Daten von Prachov dazu, dass sich lediglich in einem Fall ein Unterschied von kaum mehr als 0.1% ergibt (was gegebenenfalls mit Rundungsfehlern o.ä. zu erklären wäre). Als ein wesentliches Ergebnis dieser vergleichenden Gegenüberstellung äußern Andreeva et al. (1965: 49) die Vermutung, in den Daten zeige sich „eine gute Übereinstimmung zwischen der Subsprache der Radioelektronik und der wissenschaftlich-technischen Sprache im allgemeinen“, während bei Fachtexten oder bei gemischten Texten eine andere Verteilung der Grapheme zu beobachten sei.

Eine statistische Überprüfung, ob die von den Autoren vorgebrachte Behauptung bezüglich der unterschiedlichen Verteilung zutrifft, wird allerdings nicht durchgeführt.

Tab. 7: Zusammenschau verschiedener Graphemuntersuchungen von Andreeva et al. (1965)

	GML (1961)	**A (1959)**	**LG (1958)**	**BF (1961)**
О	0,0955	0,0940	0,0900	0,0910
Е	0,0765	0,0800	0,0720	0,0725
И	0,0740	0,0680	0,0620	0,0610
Т	0,0610	0,0580	0,0530	0,0545
Н	0,0605	0,0600	0,0530	0,0630
А	0,0600	0,0590	0,0620	0,0705
Р	0,0445	0,0450	0,0400	0,0510
С	0,0440	0,0470	0,0450	0,0405
В	0,0360	0,0380	0,0380	0,0495
Л	0,0355	0,0370	0,0350	0,0215
П	0,0275	0,0270	0,0230	0,0320
К	0,0270	0,0250	0,0280	0,0230
М	0,0270	0,0330	0,0260	0,0290
Д	0,0245	0,0230	0,0250	0,0335
У	0,0200	0,0190	0,0210	0,0175
Я	0,0190	0,0180	0,0180	0,0215
Ы	0,0175	0,0190	0,0160	0,0155
Ч	0,0145	0,0130	0,0120	0,0155
З	0,0140	0,0140	0,0160	0,0150
Х	0,0112	0,0100	0,0090	0,0115
Й	0,0108	0,0120	0,0100	0,0135
Ь	0,0106	0,0140	0,014*	0,0145
Б	0,0102	0,0110	0,0140	0,0150
Ю	0,0094	0,0060	0,0060	0,0052
Г	0,0092	0,0110	0,0130	0,0096
Ж	0,0057	0,0070	0,0070	0,0083
Э	0,0054	0,0050	0,0030	0,0015
Ш	0,0050	0,0050	0,0060	0,0043
Ф	0,0047	0,0040	0,0020	0,0015
Ц	0,0044	0,0050	0,0040	0,0025
Щ	0,0037	0,0050	0,0030	0,0030
Ъ	0,0003	0,0010	*	0,0003

Führt man eine entsprechende Re-Analyse der obigen Daten durch, um die von Andreeva et al. aufgestellte Behauptung zu überprüfen, dann zeigt sich allerdings, dass die von den Autoren gezogene Schlussfolgerung nicht zutrifft. Diese Überprüfung lässt sich leicht wie folgt anstellen: Da die Daten nicht in Rohform, sondern nur in Form von prozentualen Häufigkeiten angeführt sind, müssen sie zunächst in Rangwerte umkodiert werden.[9] Damit ergibt sich die Möglichkeit, mit Hilfe von nicht-parametrischen Tests zu überprüfen, ob zwischen den mittleren Rängen der verbundenen Stichproben signifikante Unterschiede bestehen, bzw. das Maß einer allfällig bestehenden Korrelation zwischen den Stichproben zu testen. Dazu eignen sich der sog. Friedman-Test bzw. die Berechnung des sog. Konkordanzkoeffizienten W nach Kendall. Im Ergebnis stellt sich heraus, dass bei Werten von $W = 0.001$ bzw. $\chi^2_{FG=3} = 0.084$ die Stichproben hochgradig homogen sind – somit lassen sich allfällige Schlussfolgerungen in Hinsicht auf textsortenspezifische Unterschiede der Graphemvorkommenshäufigkeit statistisch nicht absichern. Dieser Befund wirft folglich eine andere Frage auf, nämlich die nach einem allgemeinen Verteilungsmodell für (russische) Grapheme – eine Frage, der im zweiten Teil dieser Abhandlung detaillierter nachzugehen sein wird.

2.8. Kalinina (1968)

In ähnlicher Weise wie Andreeva et al. (1965) hat auch Kalinina (1968: 82) die Ergebnisse zweier verschiedener Stichproben vergleichend gegenübergestellt: zum einen die absoluten Vorkommenshäufigkeiten auf der Basis von 15.620 Wortformen (100.000 Grapheme) aus dem Bereich der Elektrotechnik, zum anderen – unter Verweis auf die Arbeiten von Jaglom/Jaglom (1960) und Charkevič (1955) die relativen Vorkommenshäufigkeiten von Graphemen in nicht näher bestimmten Texten der russischen Literatur – wobei nach dem oben Dargestellten offensichtlich ist, dass es sich de facto um die Daten von Prachov (1946) handelt. Tab. 8 enthält die entsprechenden Daten.

Im Gegensatz zu Andreeva et al. (1965) hat Kalinina (1968) versucht, die Beziehung der Vorkommenshäufigkeit der Grapheme in beiden Stichproben statistisch zu testen.

[9] Im gegebenen Fall können entsprechende Tests nur für 30 Grapheme durchgeführt werden, da in einer der Stichproben (LG 1958) die beiden Grapheme „ь“ und „ъ“ nicht differenziert wurden.

Tab. 8: Graphemhäufigkeiten nach Kalinina (1968)

	Elektrotechnik			Literatur	
Graphem	**f(abs)**	**p**	**Rang**	**p**	**Rang**
О	11376	0,11376	1	0,110	1
Е	8907	0,08907	2	0,087	2
И	7852	0,07852	3	0,075	4
Т	7338	0,07338	4	0,065	5
А	7020	0,0702	5	0,075	3
Н	6889	0,06889	6	0,065	6
Р	5498	0,05498	7	0,048	8
С	5116	0,05116	8	0,055	7
Л	4227	0,04227	9	0,042	10
В	4104	0,04104	10	0,046	9
К	3358	0,03358	11	0,034	11
П	3072	0,03072	12	0,028	14
М	3047	0,03047	13	0,031	12
Д	2641	0,02641	14	0,03	13
Я	2302	0,02302	15	0,022	16
Ы	1919	0,01919	16	0,019	17
У	1915	0,01915	17	0,025	15
Ч	1752	0,01752	18	0,015	22
З	1563	0,01563	19	0,018	18
Ь, Ъ	1364	0,01364	20	0,017	19
Г	1256	0,01256	21	0,016	21
Б	1210	0,0121	22	0,017	20
Х	1200	0,0120	23	0,011	24
Й	1032	0,01032	24	0,012	23
Э	789	0,00789	25	0,003	30
Ж	753	0,00753	26	0,009	25
Ю	692	0,00692	27	0,007	26
Ц	477	0,00477	28	0,005	28
Щ	460	0,0046	29	0,004	29
Ф	449	0,00449	30	0,002	31
Ш	422	0,00422	31	0,007	27

Da in der einen Stichprobe absolute, in der anderen relative Häufigkeiten vorliegen, hat sie zunächst diese Häufigkeiten in Ränge transformiert. In weiterer Folge verwendete sie unter Bezugnahme auf eine frühe (1925 auch ins Deutsche, 1939 ins Englische übersetzte) Arbeit zur Korrelationstheorie von A.A. Čuprov (1925) die folgende Formel (1) zur Berechnung eines Koeffizienten α, der die Stärke der Korrelation ausdrückt:

$$\alpha = 1 - \frac{3 \cdot \sum_{i=1}^{n} |i - j_i|}{n^2 - 1} \tag{1}$$

Die Summe $\sum_{i=1}^{n} |i - j_i|$ stellt hier die Summe der absoluten Differenzen zwischen den Rängen in beiden Stichproben dar, mit *n* wird der Inventarumfang bezeichnet. Bei absolut kongruenter Beziehung beträgt der Koeffizient α = 1, bei absoluter Diskongruenz liegt der Koeffizient bei 0: Im konkreten Fall beträgt α = 0.88, so dass nach Kalinina (1968) von einer starken Kongruenz zwischen beiden Stichproben auszugehen ist.

Es ist leicht zu sehen, dass der Koeffizient α dem Koeffizienten eines anderen, heute gewöhnlich verwendeten nicht-parametrischen Tests sehr ähnlich ist, nämlich dem Spearman'schen Rangkorrelationskoeffizienten ρ, der nach der Formel

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} |i - j_i|^2}{n^3 - n} \tag{2}$$

zu berechnen ist. Berechnet man aufgrund der in Tab. 8 angeführten Daten ρ nachträglich, so ergibt sich im gegebenen Fall ein Wert von $\rho = 0.981$; dieser Wert lässt sich über die *t*-Verteilung auf seine Signifikanz hin prüfen, was zu dem Ergebnis führt, dass beide Stichproben als hochgradig signifikant homogen anzusehen sind ($p < 0.001$). Damit bestätigt sich auch an den Daten von Kalinina (1968) auf andere Art und Weise das Ergebnis unserer oben durchgeführten Re-Analyse von Andreeva et al. (1965), aus der ja im wesentlichen die Homogenität der vier miteinander verglichenen Stichproben hervorging. Dieser Befund wirft somit abermals die Frage nach einem allgemeinen Verteilungsmodell für (russische) Grapheme auf.

2.9. Žuravlev (1970)

Žuravlev (1970) hat im Zusammenhang mit der Frage, ob sich „lebendige" Umgangssprache von der stilisierten Rede literarischer Figuren unterscheidet, unter anderem die Graphemhäufigkeiten in entsprechenden Texten miteinander verglichen. Als Textmaterial dienten ihm drei verschiedene Textgruppen:

I. Tonbandaufzeichnungen von Gesprächen zwischen Lehrern und Studenten («RR_1»),
II. Auswertungen von Gesprächsaufzeichnungen, die ursprünglich von Peškovskij (1925) stammen («RR_2»); dabei hat Žuravlev allerdings ganz offensichtlich nicht die von Peškovskij (1925) selbst angegebenen Laut-Häufigkeiten übernommen, sondern die entsprechenden Graphem-Häufigkeiten neu berechnet.
III. Redepassagen aus verschiedenen Texten der sowjetischen Literatur («SR»).

Als Basis für diese drei Textgruppen dienten jeweils 10 Zufallsstichproben à 1000 Zeichen; gezählt wurde für jede Stichprobe die Vorkommenshäufigkeit aller Buchstaben des Alphabets. Žuravlevs (1970) zentrale Frage lautete, inwiefern die Vorkommenshäufigkeit der einzelnen Grapheme (a) von Teilstichprobe zu Teilstichprobe und (b) zwischen den drei Textgruppen stabil ist bzw. inwiefern zu beobachtende Schwankungen als zufällige Abweichungen anzusehen sind; dann wären die entsprechenden Texte bzw. Textgruppen als (im Hinblick auf die untersuchte Frage) homogen, andernfalls als heterogen anzusehen.

Unter Bezugnahme auf einschlägige Überlegungen bei Golovin (1966: 25ff) ging Žuravlev wie folgt vor:

1. als erstes wurde innerhalb jeder der 30 Teilstichproben die Vorkommenshäufigkeit der einzelnen Grapheme berechnet;
2. dann wurde für die jeweils zehn Teilstichproben der drei Textgruppen (RR_1, RR_2, SR) die mittlere Vorkommenshäufigkeit aller Grapheme berechnet, so dass sich für jeden Buchstaben drei Mittelwerte (einer für jede Textgruppe) ergaben: $\bar{x}_{RR1}$, $\bar{x}_{RR2}$, $\bar{x}_{SR}$.
3. von jeder Vorkommenshäufigkeit aller Buchstaben (x_i) in den 30 Teilstichproben wurde der jeweilige Gruppenmittelwert (also $\bar{x}_{RR1}$, $\bar{x}_{RR2}$, oder $\bar{x}_{SR}$) subtrahiert, dann wurde innerhalb jeder der drei Textgruppen die Summe der quadrierten Differenzen durch den jeweiligen Gruppenmittelwert dividiert.

Aufgrund dieser Schritte ergab sich so gemäß der Formel

$$\chi^2 = \frac{\sum\left(x_i - \bar{x}\right)^2}{\bar{x}} \tag{3}$$

ein χ^2-Wert als Maß der Homogenität. Die Bewertung über Vorliegen von Homogenität hängt natürlich von der eingeräumten Fehlerwahrscheinlichkeit ab, die traditionell bei 5% oder 1% festgelegt wird. In Abhängigkeit von dieser Fehlerwahrscheinlichkeit liegt der Schrankenwert von χ^2 bei $n - 1 = 9$ Freiheitsgraden bei 5%-iger Fehlerwahrscheinlichkeit bei $\chi^2 = 16.92$, bei 1%-iger Fehlerwahrscheinlichkeit bei $\chi^2 = 21.67$. Tab. 9 stellt die erhaltenen Ergebnisse dar: Mit $\bar{x}_{RR1}$, $\bar{x}_{RR2}$, bzw. $\bar{x}_{SR}$ sind für die einzelnen Grapheme die jeweiligen, sich aus jeweils zehn Teilstichproben ergebenden Gruppenmittelwerte bezeichnet.

Bei der Interpretation dieser Ergebnisse ergibt sich Folgendes: Wenn man eine 1%-ige Fehlerwahrscheinlichkeit zulässt, so ist leicht zu sehen, dass mit Ausnahme des Graphems “У” in der Textgruppe der stilisierten Rede (*SR*) die Vorkommenshäufigkeit aller Grapheme in allen Teilstichproben eine hohe Homogenität aufweist: Nur in diesem einen Fall ist $\chi^2 > 21.67$. Demnach wäre die Schwankung der Vorkommenshäufigkeit der Grapheme zwischen den einzelnen Stichproben also unabhängig von der Textgruppe als extrem gering anzusehen.

Žuravlev lässt allerdings in Anlehnung an Golovin eine 5%-ige Fehlerwahrscheinlichkeit zu; unter dieser Bedingung gelangt er zu einer etwas anderen Schlussfolgerung: In diesem Fall tendiert nämlich die Vorkommenshäufigkeit der einzelnen Grapheme in den beiden umgangssprachlichen Stichproben (RR_1 und RR_2) eher zur Homogenität, während die stilisierte Rede (*SR*) in dieser Hinsicht stärker schwankt – immerhin liegen in dieser Textgruppe 13 der 33 Werte über dem χ^2-Wert von 16.92.

Dieser Umstand würde also so zu interpretieren sein, dass die zehn Teilstichproben der beiden umgangssprachlichen Textgruppen in sich stark homogen sind, während bei den Texten der stilisierten Rede jeweils eigene, deutlich von den anderen Teilstichproben derselben Textgruppe verschiedene Buchstabenhäufigkeiten vorliegen. Dieser Umstand ließe sich gegebenenfalls mit dem unterschiedlichen Ausmaß der Heterogenität des zugrundeliegenden Materials der Teilstichproben erklären: Während Žuravlev bei der Analyse der stilisierten Rede (*SR*) nämlich vollkommen unterschiedliche Texte von verschiedenen Autoren heranzog, basierte die Analyse der Umgangssprache (RR_1) auf einem einheitlichen Textmaterial, ebenso wie (weitestgehend) auch die in 1000-er Teilstichproben zerlegten Gesprächsaufzeichnungen von Peškovskij (1925).

Tab. 9: Homogenität von Graphemhäufigkeiten in verschiedenen Textgruppen (nach Žuravlev 1970)

		RR1		***RR2***		***SR***	
No.	**Graphem**	$\bar{x}_{RR1}$	χ^2	$\bar{x}_{RR2}$	χ^2	$\bar{x}_{SR}$	χ^2
1	а	97	14	93	20	88	12
2	б	18	15	18	8	17	16
3	в	36	16	39	14	41	14
4	г	15	12	14	4	16	10
5	д	39	20	34	12	34	19
6	е	83	9	82	16	88	18
7	е	8	10	7	4	8	17
8	ж	10	6	7	6	12	9
9	з	13	5	14	10	17	7
10	и	54	9	57	14	57	20
11	й	13	12	11	8	10	7
12	к	36	13	30	12	35	10
13	л	36	11	39	12	34	10
14	м	30	10	31	18	40	20
15	н	63	13	61	16	64	14
16	о	104	8	105	12	107	14
17	п	26	6	27	10	22	10
18	р	36	5	38	8	39	20
19	с	50	6	44	14	45	21
20	т	75	10	76	24	69	6
21	у	31	9	29	16	35	27
22	ф	2	8	2	4	1	21
28	х	7	12	10	6	7	19
24	ц	4	5	3	4	3	7
25	ч	20	10	23	6	20	18
26	ш	12	14	10	4	13	15
27	щ	3	8	3	2	4	21
28	ъ	0	0	0	0	0	0
29	ы	16	17	16	10	17	18
30	ь	25	8	24	8	25	8
31	э	3	7	5	4	5	12
32	ю	6	5	8	8	8	8
33	я	21	5	24	8	24	14

2.10. Grigor'ev (1980)

Grigor'ev (1980a, b) hat die Vorkommenshäufigkeit russischer Grapheme untersucht und in zwei verschiedenen Darstellungen behandelt. Im Hinblick auf das Untersuchungsmaterial spricht Grigor'ev (1980a) in der ersten der beiden Darstellungen von einem nicht näher spezifizierten „Textfragment der gegenwärtigen künstlerischen Prosa" mit einem Gesamtumfang von insgesamt 50.000 Graphemen; in der zweiten der beiden Darstellungen spricht Grigor'ev (1980b: 43) von „three coherent parts of a Russian text […] representing the initial part of a novel", die er untersucht habe, und deren Umfang sich auf 100.000 Grapheme belaufe. Tab. 10 stellt die Ergebnisse beider Untersuchungen vergleichend nebeneinander.[10]

In beiden Untersuchungen ging es Grigor'ev vornehmlich darum, die Güte der Ergebnisse kleinerer Stichproben im Vergleich zu größeren Stichproben bzw. zur Gesamtstichprobe zu bewerten. Aus diesem Grunde hat er die beiden Gesamtstichproben jeweils in bestimmte „Portionen" von 100, 500 oder 1000 Graphemvorkommnissen unterteilt, und dann entweder die Ergebnisse dieser Portionen miteinander verglichen oder aber der Gesamtstichprobe vergleichend gegenübergestellt. In der ersten der beiden Studien hat Grigor'ev (1980a) zum Zwecke dieses Vergleichs den Spearman'schen Rangkorrelationkoeffizienten ρ berechnet (s.o., Formel 2); dabei hat er festgestellt, dass ρ bei Vergleichen zwischen den kleineren Teilstichproben jeweils zwischen ca. $\rho = .64$ bis $\rho = .79$ lag, dass er bei den größeren Teilstichproben allerdings Werte zwischen $\rho = .98$ bis $\rho = .99$ annahm. Die daraus gezogene Schlussfolgerung einer deutlichen Homogenität bei den größeren Teilstichproben bestätigte sich nach Grigor'ev (1980b) auch bei einem Vergleich von Teilstichproben im Umfang von 10.000, 50.000 bzw. 100.000 Graphemvorkommnissen; in dieser zweiten Studie führte Grigor'ev zur Überprüfung auf Homogenität Kolmogorov-Smirnov-Tests durch, und schloss aufgrund der nach $\lambda = D_{\max}\sqrt{n}$ berechneten Werte ($\lambda_1 = 0.541$, $\lambda_2 = 0.367$, $\lambda_3 = 0.322$) auf eine „increasing conformity of distribution".

[10] In der ersten Darstellung wird das Graphem „ë" nicht als eigenes Graphem tabelliert; es wird nur in einer Fußnote eine Vorkommenshäufigkeit von 0,00016 erwähnt, in der zweiten Darstellung wird von insgesamt 14 Vorkommnissen (d.h. 0,00014%) gesprochen. Das erklärt, warum sich die hier angegebenen Werte der Arbeit von 1980b nicht auf 100.000, sondern auf 99.986 belaufen.

Tab. 10: Graphemhäufigkeiten nach Grigor'ev (1980a,b)

		1980a		1980b	
Rang	**Graphem**	**f(abs)**	**f(%)**	**f(abs)**	**f(%)**
1	О	5678	11,36	11410	11,41
2	Е	4206	8,41	8610	8,61
3	А	3979	7,96	8002	8,00
4	И	3349	6,70	6536	6,54
5	Н	3112	6,22	6097	6,10
6	Т	2983	5,97	5926	5,93
7	С	2511	5,02	5072	5,07
8	Л	2334	4,67	4674	4,67
9	В	2174	4,35	4492	4,49
10	Р	2091	4,18	4140	4,14
11	К	1981	3,96	4157	4,16
12	М	1555	3,11	3095	3,10
13	У	1527	3,05	3098	3,10
14	Д	1493	2,99	2977	2,98
15	П	1294	2,59	2488	2,49
16	Ы	1068	2,14	2090	2,09
17	Я	1052	2,10	2092	2,09
18	Б	990	1,98	1981	1,98
19	Ь	968	1,94	1939	1,94
20	Г	923	1,85	1912	1,91
21	Ч	798	1,60	1611	1,61
22	З	780	1,56	1490	1,49
23	Й	706	1,41	1373	1.37
24	Ж	557	1,11	1130	1,13
25	Х	512	1,02	1012	1,01
26	Ш	480	0,96	857	8,57
27	Ю	341	0,68	685	6,85
28	Щ	170	0,34	323	3,23
29	Э	168	0,34	310	3,10
30	Ц	162	0,32	304	2,04
31	Ф	42	0,08	81	0,08
32	Ъ	16	0,03	22	0,02

Eine solche Interpretation der Ergebnisse liegt zwar nahe, ist aber bei beiden angewendeten Verfahren insofern problematisch, als bei den größeren Stich-

proben die jeweils kleineren als Teilmenge inkludiert waren: so waren z.B. die 50.000 Grapheme Bestandteil der 100.000er Gesamtstichprobe, usw.). Insofern haben wir es nicht mit unabhängigen Stichproben zu tun – der Kolmogorov-Smirnov-Test ist jedoch explizit nur für unabhängige Stichproben zulässig, so dass die Berechtigung der Schlussfolgerung in Frage zu stellen ist.

Wenn man also die Schlussfolgerung des Autors überprüfen will, muss man anders vorgehen – und dann lässt sie sich auch untermauern: Da die absolute Vorkommenshäufigkeit der ersten 50.000 Grapheme in Grigor'ev (1980a) und die der gesamten 100.000 Grapheme in Grigor'ev (1980b) angegeben ist, kann man ohne weiteres von den Vorkommenshäufigkeiten in der Gesamtstichprobe diejenige der 50.000er-Stichprobe subtrahieren; so erhält man zwei unabhängige Stichproben im Umfang von jeweils 50.000. In diesem Fall weist der Spearman'sche Rangkorrelationkoeffizient $\rho = .998$ den Zusammenhang zwischen den beiden Stichproben als hoch signifikant aus ($p < .001$); auch ein mit diesen beiden Datensätzen durchgeführter Kolmogorov-Smirnov-Test zeigt in der Tat die hohe Homogenität beider Stichproben auf ($z = .375$, $p = .999$). Abb. 1 veranschaulicht den nahezu perfekten linearen Zusammenhang zwischen den beiden Stichproben.

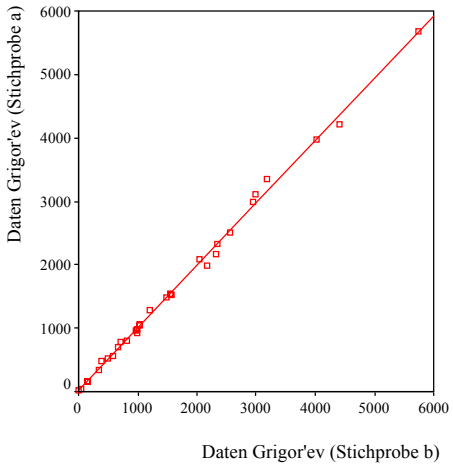

Abb. 1: Zusammenhang zwischen den Stichproben

Damit werfen auch die Untersuchungen von Grigor'ev die Frage nach einem einheitlichen Modell für die Vorkommenshäufigkeit (russischer) Grapheme auf.

2.11. Dietze (1982)

Die letztendlich auch bei Grigor'ev im Vordergrund stehende Leitfrage ist offenbar die nach der Zuverlässigkeit der Ergebnisse in Abhängigkeit von der Stichprobengröße und dem zugrunde liegenden Textmaterial. Dieser Frage ist Dietze (1982) in einer eigenen Untersuchung auf andere Art und Weise nachgegangen. In dieser Untersuchung wurde die Graphemhäufigkeit in russischen fachsprachlichen Texten untersucht; als Material dienten 500 Referate aus der sowjetischen Referatezeitschrift *Referativnyj žurnal* zum Thema „wissenschaftliche und technische Information". Das so zusammengestellte Korpus beinhaltete insgesamt 57.666 Wortformen mit insgesamt 429.257 Graphemen, wobei der Unterschied zwischen Groß- und Kleinbuchstaben vernachlässigt und das Graphem „ë" als „e" ausgewertet wurde.

Tab. 11 enthält die in absteigender Rangreihenfolge sortierten absoluten und prozentualen Vorkommenshäufigkeiten der 32 untersuchten Grapheme.

Tab. 11: Vorkommenshäufigkeit russischer Grapheme nach Dietze (1982)

Rang	Graphem	F(abs)	f(%)	Rang	Graphem	f(abs)	f(%)
1	О	44172	10,29	17	У	8413	1,96
2	И	42024	9,79	18	З	7000	1,63
3	Е	35662	8,31	19	Б	6464	1,51
4	А	33967	7,91	20	Ч	6005	1,40
5	Н	29877	6,96	21	Х	5390	1,26
6	Т	27447	6,39	22	Й	4852	1,13
7	С	26034	6,06	23	Г	4716	1,10
8	Р	22279	5,19	24	Ц	4491	1,05
9	В	17586	4,10	25	Ь	4389	1,02
10	Л	14613	3,40	26	Ф	3912	0,91
11	К	14189	3,31	27	Ю	2904	0,68
12	М	13890	3,24	28	Ж	2537	0,59
13	П	12736	2,97	29	Щ	1670	0,39
14	Д	11079	2,58	30	Ш	1224	0,29
15	Я	9893	2,30	31	Э	1054	0,25
16	Ы	8632	2,01	32	Ъ	156	0,04

Dietze (1982: 81) ging es, wie gesagt, nicht zuletzt um die Frage, „ob eine Stichprobe für sprachstatistische Forschungen groß genug ist". Um dieser Frage nachzugehen, bezieht er sich zunächst auf die folgende, von Frumkina

(1973: 282) angeführte Formel[11], mit der sich der sog. „relative Fehler" einer Stichprobe, d.h. die Abweichung in Bezug auf die relative Häufigkeit, berechnen lässt:

$$\delta = \frac{z_\rho}{\sqrt{Np}} \tag{4}$$

Hierbei ist δ der relative Fehler, N ist der Stichprobenumfang (also im gegebenen Beispiel die Gesamtzahl der berechneten Buchstaben), p ist die (relative) Häufigkeit der zu untersuchenden Einheit, z_ρ ist eine Konstante, die von der gewählten Konfidenz der Schätzung abhängt (üblicherweise bezieht man sich hierbei auf das 95%-Quantil der Normalverteilung, was einem z-Wert von 1.96 entspricht; bei 99% beträgt $z = 2.576$).

Wenn man nun – wie Dietze (1982: 82) – das arithmetische Mittel aller berechneten Grapheme als p ansetzt, das im konkreten Fall $\bar{x} = 13414.28$ beträgt, und in die obige Formel (4) einsetzt, erhält man einen mittleren relativen Fehler von 0.0169 (d.h. ca. 1.7%); für die Vorkommenshäufigkeit des am seltensten vorkommenden Graphems (das ist ь mit $f_i = 156$) erhält man einen relativen Fehler von 0.1569 (d.h. ca. 15.7%). Beide Werte hält Dietze (1982: 82) für „statistisch repräsentativ".

Interessant ist, dass sich aufgrund des oben gezeigten Ansatzes unter Festlegung einer zulässigen relativen Abweichung (in der Regel nimmt man hier 10%) und unter Festlegung der Konfidenz die notwendige Stichprobengröße bestimmen lässt (was Dietze allerdings nicht tut). Auch eine solche Formel ist bei Frumkina (1973: 286) im weiteren Verlauf ihrer Überlegungen angeführt. Demnach ergäbe sich aus

$$0.10 = \frac{1.96}{\sqrt{Np}}$$

für die mittlere Vorkommenshäufigkeit eine Stichprobengröße von $N = 12.293$; bei der notwendigen Ausrichtung am seltensten Vorkommnis würde die Stichprobengröße bereits $N = 1057073$ betragen.

Allerdings ist dazu zu sagen, dass beide Formeln von Frumkina einen Fehler aufweisen, so dass die soeben angestellten Überlegungen im Prinzip zwar in die richtige Richtung zielen, die entsprechenden Berechnungen allerdings theoretisch falsch bzw. praktisch sehr stark approximiert sind. So gilt für die Berechnung des relativen Fehlers nicht die oben dargestellte, sondern die folgende Formel:

[11] Dietze notiert diese Formel allerdings falsch, da er die Wurzel im Nenner nicht anführt.

(5) $$\delta = \frac{z_\rho \sqrt{q}}{\sqrt{Np}}$$

Dietze hat also – in Anlehnung an Frumkina – das als *1-p* zu berechnende *q* im Zähler ausgelassen, was im gegebenen Fall allerdings zu nur geringfügig verschiedenen Werten von relativen Fehlern von $\delta = 0.0167$ für den mittleren relativen Fehler führt (beim seltensten Graphem ändert sich so gut wie nichts). Damit verändert sich natürlich auch die Berechnung der Stichprobengröße; so lässt sich aus der Formel (1) die Berechnung für *N* nach

$$N = \frac{z_{\alpha/2}^2 q}{\delta^2 p}$$

umordnen (wobei man $z_{\alpha/2}$ wegen der zweiseitigen Abweichung benötigt). Auch hier unterscheiden sich die Werte in der Praxis nur geringfügig: bezieht man sich auf das arithmetische Mittel, so ergibt sich mit $z_{\alpha/2} = 1.96$ und $\delta = 0.1$ eine notwendige Stichprobengröße von 11909, aus dem kleinsten Wert ergibt sich $N = 1056689$.

Abgesehen von der fehlerhaften Berechnung einer „repräsentativen" Stichprobengröße, die sich bei Bezugnahme auf Frumkina ergäbe, ist die von Dietze vorgeschlagene Bezugnahme auf das arithmetische Mittel allerdings vor allem deswegen problematisch, weil die Graphemhäufigkeiten seiner Stichprobe nicht die Voraussetzung der Normalverteilung erfüllen – das zeigt ein Kolmogorov-Smirnov-Test auf Normalverteilung: bei einem Wert von 0.181 ist die Abweichung von der Normalverteilung hochsignifikant ($p < 0.01$). Eine Bezugnahme auf den kleinsten Wert hingegen hätte einen anderen Nachteil: in diesem Fall nämlich würden alle anderen Werte vollständig außer Acht bleiben.

Insofern wäre aus heutiger Sicht ein Vorgehen wie etwa das von Kubáček (1994) vorgeschlagene vorzuziehen. Mit diesem Verfahren lässt sich der notwendige Stichprobenumfang gleichzeitig über alle Grapheme schätzen; dieser Ansatz geht von der mittleren Standardabweichung *r* aller relativen Häufigkeiten aus, die gegeben ist als

(6) $$r = \frac{1}{\sqrt{N}} \prod_{i=1}^{K} p_i^{\frac{1}{2(K-1)}},$$

wobei *K* der Inventarumfang und *N* der Stichprobenumfang ist. Die einzelnen p_p schätzt man aus der Stichprobe als $\hat{p}_i = f_i / N$. Löst man (6) für *N* auf, so bekommt man

(7) $$N = \frac{1}{r^2} \prod_{i=1}^{K} \hat{p}_i^{\frac{1}{K-1}}$$

Die Berechnung führt man in der Praxis am besten logarithmisch durch, d.h. als

$$\ln N = \frac{1}{K-1}\sum_{i=1}^{K} \ln p_i - 2\ln r \tag{8}$$

Da im oben gegebenen Fall $\sum_{i=1}^{32} \ln p_i = -128.26$, erhält man für r = 0.001 nach (8)

$$\ln N = -128.26/(32-1) - 2 \cdot \ln(0.001) = 9.678$$

und damit einen geschätzten Stichprobenumfang von

$$\widehat{N} = e^{9.678} = 15965.65$$

Für r = 0.0005 betrüge die Stichprobengröße entsprechend $\widehat{N} = 63862.61$

3. Resümee und Perspektiven

Versuchen wir, abschließend zu einem Resümee zu gelangen. Es sollte aus den vorangegangenen Darstellungen deutlich geworden sein, dass das „simple Zählen" von Buchstaben – das hier am Beispiel des Russischen veranschaulicht wurde – niemals nur Selbstzweck war: Immer ging es um weiterführende Fragen, angefangen von mathematischen und methodologischen Problemen, über Fragen der Optimierung technischer Einrichtungen oder der Strukturierung von Codes und Prozessen der Informationsübertragung, bis hin zu Fragen der Textstilistik und Texttypologie. Abgesehen von der Unterschiedlichkeit der betreffenden Fragen hat sich auch gezeigt, dass mit den jeweiligen Untersuchungen unterschiedliche Herangehensweisen verbunden waren, um die Spezifik oder aber übergreifende Charakteristik der Häufigkeitsverteilung (russischer) Grapheme genauer bestimmen zu können.

Auch die vorliegende Darstellung, das wurde eingangs hervorgehoben, versteht sich ja nur als ein erster Schritt in einer Reihe konsekutiver Analysen: Die nächsten Schritte – das liegt aufgrund der Befunde der hier dargestellten Untersuchungen nahe – müssen auf die Frage eines allgemeinen Modells zielen. Dazu wird es in einem ersten Schritt notwendig sein, die in der bisherigen Forschung zur Diskussion gestellten Verteilungsmodelle zu diskutieren, um davon ausgehend allfällige theoretische Zusammenhänge zwischen diesen zu prüfen, und um sodann die zur Diskussion stehenden Model-

le auf ihre Eignung für russische Graphemhäufigkeiten zu testen. Erst im Anschluss daran wird es sinnvoll sein, weitere Sprachen in Betracht zu ziehen, um schließlich theoretische Kenngrößen wie die theoretische Entropie oder die theoretische Wiederholungsrate ableiten zu können.

Literatur

Altmann, G. (1972): „Status und Ziele der quantitativen Sprachwissenschaft.“ In: Jäger, S. (Hrsg.), *Linguistik und Statistik.* Braunschweig, 1-9.

Altmann, G. (1973): „Mathematische Linguistik.“ In: Koch, W. A. (Hrsg.), *Perspektiven der Linguistik.* Stuttgart, 208-232.

Andreeva, L. D.; Kordi, E. E.; Smirnova, L. N.; Fedulova, N. I.; Fitialova, I. B.; Fichman, B. S: (1965): „Polučenie pervogo morfologičeskogo tipa russkogo jazyka v pod“ jazyke radioėlektroniki posredstvom algoritma statistiko-kombinatornogo modelirovanija.“ In: *Statistiko-kombinatornoe modelirovanie jazykov.* Moskva, 49-64.

Bátori, I. S.; Lenders, W.; Putschke, W. (eds.) (1989): *Computational Linguistics. Computerlinguistik*. Berlin: W. de Gruyter, 113-119.

Belonogov, G. G.; Frolov, G. D. (1963): „Ėmpiričeskie dannye o raspredelenii bukv v russkoj pis’mennoj reči“, in: *Problemy kibernetiki, 3*; 287-305.

Budilovič, A. S. (1883): *Načertanie cerkovno-slavjanskoj grammatiki primenitel’no k obščej teorii russkago i drugich rodstvennych jazykov*. Varšava.

Charkevič, A. A. (1955): *Očerki obščej teorii svjazi*. Moskva.

Čuprov, A. A. (1925): *Osnovnye problemy teorii korreljacii*. Moskva, 1960. [Deutsch: *Grundbegriffe und Grundprobleme der Korrelationstheorie*. Leipzig etc., 1925. – Englisch: *Principles of the mathematical theory of correlation*. London etc., 1939.]

Dietze, J. (1982): „Grapheme und Graphemkombinationen der russischen Fachsprache.“ In: Lehfeldt, W.; Strauss, U. (eds.), *Glottometrika 4*. Bochum, 80-94.

Förstemann, E. (1846): „Über die numerischen Lautverhältnisse im Deutschen“, in: *Germania. Hrsg. von der Berlinischen Gesellschaft für deutsche Sprache und Alterthumskunde, Bd. 7*; 83-90.

Förstemann, E. (1852): „Numerische Lautverhältnisse im Griechischen, lateinischen und Deutschen“, in: *Zeitschrift für Vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen, 1*; 163-179.

Frumkina, R. M. (1973): „Zur Anwendung statistischer Methoden in der Sprachforschung.“ In: *Sprachstatistik.* Berlin (DDR), 272-298.

Gladkij, A. V.; Mel'čuk, I. A: (1969): Ėlementy matematičeskoj lingvistiki. Moskva. [Deutsch: *Elemente der mathematischen Linguistik*. Berlin, 1973.]

Golovin, B. N. (1966): *Iz kursa lekcij po lingvističeskoj statistike*. Gor'kij.

Grigor'ev, V. I. (1980a): „O dinamike raspredelenij bukv v tekste". In: *Aktual'nye voprosy strukturnoj i prikladnoj lingvistiki. Sbornik statej.* Moskva, 40-48.

Grigor'ev, V. I. (1980b): „Frequency distribution of letters and their ranks in a running text." In: *Symposium: Computational Linguistics and Related Topics*. *Summaries.* Tallinn, 43-47.

Grzybek, P.; Kelih, E.; Altmann, G. (2004): „Graphemhäufigkeiten. Teil II: Theoretische Modelle der Häufigkeitsverteilung (mit einer empirischen Untersuchung russischer Graphemhäufigkeiten)", in: *Anzeiger für Slavische Philologie*, XXXII.

Jaglom, A. M.; Jaglom, I. M. (1960): *Verojatnost' i informacija*. Moskva.

Kalinina, E. A. (1968): „Izučenie leksiko-statističeskich zakonomernostej na osnove vjerojatnostnoj modeli". In: *Statistika reči*. Leningrad, 64-107.

Kubáček, L. (1994): „Confidence limits for proportions of linguistic entities", in: *Journal of Quantitative Linguistics 1;* 56-61.

Lebedev, D. S.; Garmaš, V. A. (1958): „O vozmožnosti uveličenija skorosti peredači telegrafnych soobščenij", in: *Ėletrosvjaz'*, 1; 68-69.

Lebedev, D. S.; Garmaš, V. A. (1959): „Statističeskij analiz trechbukvennych sočetanij russkogo teksta." In: *Problemy peredači informacii, vyp. 2*. Moskva, 78-80.

Lenz, S. (2000): *Korpuslinguistik*. Tübingen.

Markov, A. A. (1913): „Primer statističeskogo izsledovanija nad tekstom »Evgenija Onegina« illjustrirujuščij svjaz' ispytanij v cep'", in: *Izvestija Imperatorskij Akademii Nauk // Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg, ser. VI, t. 7, no. 3*; 153-162.

Morozov, N. A. (1915): „Lingvističeskie spektry", in: *Izvestija otdelenija russkago jazyka i slovesnosti imperatorjskoj akademii nauk, XX(1-4)*; 95-127.

Ol'chin, P. (1907): „Pervaja opora pri postroenii racional'noj stenografii", in: *Stenograf. Ežemesjačnyj žurnal, posvjaščennyj voprosam naučnoj i praktičeskoj stenografii*, *no. 4-5*; 114-118.

Peškovskij, A. M. (1925): „Desjat tysjač zvukov." In: Dsb., *Metodika rodnogo jazyka, lingvistika, stilistika, poėtika*. Leningrad/Moskva, 167-191.

Prachov, P. V. (1946): *Sravnenie startstopnych telegrafnych apparatov s četyrechrjadnoj i trechrjadnoj klaviaturoj*. Moskva. [= Otčet CNIIS.]

Proskurnin, N. (1933): „Podsčety častoty liter i komplektovka šrifta." In: *Revoljucija i pis'mennost'. Sbornik I*. Moskva-Leningrad, 72-82.

Schleicher, A. (1852): *Die Formenlehre der kirchenslawischen Sprache, erklärend und vergleichend dargestellt*. Bonn u.a.
Spraul, H.(1999): „Graphemik." In: Jachnow, H. (ed.), *Handbuch der sprachwissenschaftlichen Russistik*. Wiesbaden, 66-86.
Uspenskij, V. A. (2002): *Trudy po nematematike. Tom 1 & 2*. Moskva.
Žuravlev, A. P. (1970): „O nekotorych otličijach živoj razgovornoj reči ot stilizovannoj." In: *Russkaja razgovornaja reč'*. Saratov, 176-184.

peter.grzybek@uni-graz.at
keli@gewi.kfunigraz.ac.at