DOI: 10.36208/AnzSlPh47201905

Anzeiger für Slavische Philologie
(XLVII) 2019, 99-121

Wörterbuch vs. Text: Häufigkeiten von Graphemen im Slowenischen (zum 130. Geburtstag von N.S. Trubeckoj)

Emmerich Kelih (Wien)

Abstract
The present contribution deals with grapheme frequencies in Slovene. The focus is on factors influencing the results of counting of grapheme frequencies in a dictionary and in a text. One further focus is the statistical modelling of grapheme frequencies on these two linguistic levels. Whereas for both, the dictionary and text one particular statistical model is appropriate the analysis of the vowel/consonant proportion and the repeat rate yields remarkable differences. In conclusion, based on the used data, no clear-cut (statistical) difference between "dictionary" and "text" can be observed.

Keywords: Slovene, grapheme frequencies, dictionary, text, vowel proportion, consonant proportion, repeat rate, N.S. Trubeckoj

1. Einleitung

In diesem Beitrag geht es um die Vorkommenshäufigkeit von Graphemen im Slowenischen. Die grundlegende Problematik geht zurück auf eine ältere, aber zentrale Überlegung des bekannten Phonologen N.S. Trubeckoj (1890-1938). Dieser hatte in seinem Werk Grundzüge der Phonologie (Trubetzkoy 1939, zitiert nach Auflage 1989) u.a. darauf verwiesen, dass bei der Untersuchung von Phonemhäufigkeiten unbedingt zwischen Zählungen in Wörterbüchern und in zusammenhängenden Texten zu unterscheiden ist. Diese Überlegung wird im folgenden Text aufgegriffen und einer empirischen Untersuchung zugeführt. Die empirische Grundlage für diese Untersuchung bilden allerdings nicht Phoneme, sondern Grapheme in einem Grund- und Aufbauwortschatz des Slowenischen. (Kelih/Vučajnk 2018) Dieser besteht einerseits aus entsprechenden Lemmata (= Wörterbuch-Ebene) und andererseits aus jeweiligen Beispielsätzen (= Text-Ebene), die die kontextuelle Verwendung der angeführten Lemmata vor Augen führen sollen. Insofern geht es um eine quantitative Analyse, die einerseits auf Systemebene (Wörterbuch) und andererseits auf der Textebene (Beispielsätze) durchgeführt wird. Nach einer einleitenden Darstellung der theoretischen Grundlagen dieser Untersuchung, die insbesondere auf den einschlägigen Überlegungen von N.S. Trubeckoj aufbaut, wird das

untersuchte Material aus statistischer Perspektive vorgestellt. Neben rein deskriptiven Aspekten wird sodann näher die Modellierbarkeit der erhaltenen Ranghäufigkeiten durch statistische Modelle behandelt. Als nächstes wird auf die Frage eingegangen, ob es zwischen der Vokal- bzw. Konsonantenhäufigkeit auf Wörterbuch- und Textebene statistisch signifikante Unterschiede zu beobachten gibt oder nicht. Des Weiteren wird auf das unterschiedliche Verhalten der Wiederholungsrate im Wörterbuch und im Text eingegangen. Abschließend erfolgt eine linguistische Interpretation und Zusammenfassung der erhaltenen Resultate.

Es ist einleitend zu bemerken, dass Grapheme und Phoneme klarerweise unterschiedliche linguistische Entitäten sind und beide gleichermaßen einer quantitativen Untersuchung zugeführt werden können. Im vorliegenden Fall ist zu bedenken, dass die slowenische Orthographie bzw. die slowenische Graphematik relativ nahe den jeweiligen Phonembestand wiedergeben (zur Phonem-Graphem-Korrespondenz im Slowenischen vgl. Kelih 2005). Dies ist natürlich nicht der Fall in Sprachen mit Zeichenschriften, aber auch z.B. im Englischen, wo die Graphem-Phonem-Korrespondenzen nicht sehr eindeutig sind. Üblicherweise ist die graphematische Darstellung der Phoneme in der Regel auch ahistorisch, da Veränderungen auf der Lautebene nicht unmittelbar auf die orthographische Ebene Einfluss haben. Ein weiterer Punkt ist, dass die Entität Graphem keine sprachübergreifende Entität darstellt (man denke an das Chinesische, Japanische, Altägyptische usw.), während dies bei Phonemen doch der Fall ist. Dennoch wäre auch im Slowenischen eine quantitative Analyse der Zeichen vorstellbar, die entsprechende Phoneme wiedergeben. Doch ist auch die Schriftlinguistik eine etablierte Disziplin, in der u.a. seit geraumer Zeit die Frage von Vorkommenshäufigkeiten von Zeichen ihren festen Platz hat. Somit lässt sich festhalten, dass sowohl die Analyse von Graphemen als auch Phonemen legitim erscheint, wenngleich auch die jeweilige Tiefe einer Phonem-Graphem-Korrespondenz zu beachten ist.

2. Graphemfrequenzen: Theoretische Grundlagen

In der quantitativen Linguistik gibt es ein ausgeprägtes Interesse an der Vorkommenshäufigkeit von linguistischen Einheiten und deren Modellierbarkeit durch entsprechende statistische (stetige bzw. diskrete) Modelle. Aus einer allgemein linguistischen Sicht ist die Vorkommenshäufigkeit von Graphemen auch als deskriptive Kenngröße von Interesse, d.h. man stellt sich die Frage, wie oft kommt eine bestimmte Einheit eigentlich in einer Sprache vor, wie oft kommen Vokale, wie oft Konsonanten usw. vor. Untersuchungen dieser Art geben Informationen über die Belastung einer Entität und hängen von den jeweiligen Bedürfnissen der Sprecher und Hörer (vgl. Köhler 2005) bei der Ko-

dierung ab. Sie sind auch korreliert mit der Anstrengung, die man aufbringen muss, um sie zu benutzen, was in bestimmten Bereichen (z.B. der Phonetik) auch mit der Umgebung, in der die Sprecher leben, zusammenhängen kann.

Es ist hier nicht der Ort, über die vielfältigen linguistischen Bereiche, in denen Graphemhäufigkeiten eine Rolle spielen können (von der Typographie über die funktionale Graphematik, der Informationstheorie bis hin zur Sprachtypologie), im Detail zu berichten (vgl. dazu Grzybek/Kelih/Stadlober 2009; Grzybek/Kelih 2005; Martindale et al. 1996). Im aktuellen Diskurs der quantitativen Linguistik geht es vor allem um die Frage von Einflussfaktoren auf die Vorkommenshäufigkeit von Graphemen und um mögliche Wechselbeziehungen mit weiteren Eigenschaften. Man vergleiche dazu Altmann/Lehfeldt (1980: 151f) und Grzybek/Kelih (2005), die u.a. zeigen konnten, dass der Inventarumfang, d.h. die Anzahl von konstituierenden Einheiten (sei es von Graphemen oder auch Phonemen) einen entscheidenden Einfluss auf die Form der Häufigkeitsverteilung hat.

Des Weiteren ist die Graphemfrequenz auch in einem Zusammenhang zur Graphemdistribution (bzw. Graphotaktik) zu sehen, zumal vorhandene Restriktionen in der Kombinierbarkeit ohne Zweifel einen Einfluss auf die Häufigkeit haben. Darüber hinaus konnte in Kelih (2012) gezeigt werden, dass sogar die Wortlänge einen unmittelbaren Einfluss bzw. eine Art von Rückkoppelung auf die Verteilung von Graphemen haben kann. Dies ist durch das sogenannte Menzerath'sche Gesetz (vgl. Altmann 1980 bzw. für einen aktuellen Beitrag vgl. Coloma 2015) begründbar, welches besagt, dass mit zunehmender Wortlänge die Silben von Wörtern einfacher, d.h. kürzer werden. Dies wiederum hat einen direkten Einfluss auf die Verteilung von Konsonanten und Vokalen innerhalb von Wörtern unterschiedlicher Länge, was wiederum die gesamte Häufigkeitsverteilung von Graphemen bzw. Phonemen beeinflussen kann. Darüber hinaus ist bekannt, dass man allein auf der Basis der Wortlänge unterschiedliche Textsorten bzw. Diskurstypen unterscheiden kann, was wiederum den Schluss zuließe, dass somit auch die entsprechenden Graphemhäufigkeiten unterschiedlich ausfallen müssen.

Der vorliegende Beitrag ist weniger dieser vielsprechenden Forschungsperspektive nach Wechselbeziehungen gewidmet, sondern versteht sich als die Wiederaufnahme einer älteren, aber bislang kaum systematisch untersuchten Frage nach einer adäquaten Untersuchungsbasis. Es stellt sich die Frage, ob und welche Unterschiede sich in Abhängigkeit von der Untersuchungsbasis ergeben können. In der Regel wird bei entsprechenden Untersuchungen zwischen einer *pragmatischen Häufigkeit* („Text") bzw. *systemischen Häufigkeit* („Wörterbuch") unterschieden (vgl. dazu Altmann/Lehfeldt 1980: 45, die sich dabei an die Terminologie von J. Greenberg anlehnen).

Genau diese Unterscheidung findet man aber bereits in einem im Grunde wenig beachteten Kapitel der GRUNDZÜGE DER PHONOLOGIE von N.S. Trubeckoj (Trubetzkoy 1939, zitiert nach Auflage 1962) „Zur phonologischen Statistik“ (ebda. 230-241), die seine Kenntnis der aktuellen Situation in diesem Bereich belegt. Unterschieden werden zwei Arten der Bestimmung von Phonemfrequenzen, einerseits die Zählung im Wörterbuch und andererseits die Zählung im zusammenhängenden Text. In Bezug auf die Untersuchung von Texten verweist er auf eigene Untersuchungen von unterschiedlichen Textsorten (wissenschaftlicher Text bzw. ein Märchen), wo sich stilbedingt große Unterschiede in den Proportionen einzelner Phonemklassen ergeben. Demgegenüber sind bei der Auszählung eines Wörterbuchs vor allem phonotaktische Restriktionen zu beachten, die sprachspezifisch sind und darüber hinaus positionell restringiert sein können und daher einen Einfluss auf die funktionelle Belastung der einzelnen Phoneme haben.

Es ist somit durchaus plausibel anzunehmen, dass auch die Häufigkeit von Graphemen im „Wörterbuch“ ein anderes Verhalten aufweist als in einem „Text“. Eine jede empirische Analyse impliziert eine entsprechende Operationalisierung nicht nur der zu untersuchenden Einheiten, sondern es ist auch zu klären, auf welche Art und Weise die genannte Dichotomie von „Wörterbuch vs. Text“ empirisch erfasst werden kann.

An dieser Stelle ist zu fragen, was eigentlich der Begriff Text für eine quantitative Häufigkeitsanalyse von Graphemen bedeutet. Die Abgrenzung eines operationalen Textbegriffes bringt eine Reihe von Schwierigkeiten mit sich, die an dieser Stelle nicht zu diskutieren sind. Als übliche Forschungspraxis hat sich, zumindest in Teilen der quantitativen Linguistik eingebürgert, jeweils ganze, abgeschlossene Texte[1] zu untersuchen. In jedem Fall geht es um natursprachliche Ausdrücke, die in der Regel den grammatischen Regeln einer Sprache entsprechen und ein über den Satz hinausgehendes Gebilde darstellen.

In Bezug auf die Systemebene wird in der quantitativen Phonologie vorgeschlagen, das sprachliche Material auf der Ebene eines Wörterbuches zu analysieren. Und auch dieser Aspekt lässt sich spezifizieren, indem man z.B. die Analyse der einzelnen Wörterbucheinträge bzw. Lemmata in Angriff nimmt. Ob dies eine tatsächlich adäquate Operationalisierung von „Wörterbuch“ darstellt, kann an dieser Stelle ebenfalls nicht eindeutig geklärt werden. Lemmata sind im Grunde genommen sich aus der lexikographischen Praxis ergebende

[1] Es ist nicht allein die Forschungspraxis, die das Interesse an ganzen, abgeschlossenen Texten bedingt. Vielmehr ist der Hintergrund die umfangreiche Diskussion zum Zipf’schen Gesetz, dessen Gültigkeit u.a. von der Art von Texten abhängt, in denen es zur Geltung kommt.

normierte Darstellungen von u.a. semantischer, syntaktischer und morphologischer Information eines bestimmten Wortschatzes einer bestimmten Sprache. Insofern ist das Lemma auch keine verlässliche sprachübergreifende Einheit, sondern vielmehr eine der vielen Möglichkeiten der Erfassung dieser Art von sprachlichem Material, die mit einigen Spezifika einhergeht. So würde z.B. die Auswertung von Graphemhäufigkeiten in slowenischen[2] Verben, die in der Form des Infinitivs lemmatisiert werden, vermutlich eine überdurchschnittliche Häufigkeit ausgewählter Grapheme, die eben den Infinitiv markieren, nach sich ziehen. Man vgl. *delati* (‚arbeiten'), wo *-ti* als Infinitivmarker als Beispiel für die überdurchschnittliche Häufigkeit von *t* und *i* steht. Ähnliches gilt auch für Nominalformen, wenn z.B. Feminina, wie *hiša -e* (‚Haus') in der Regel im Nominativ auf *-a* enden und deren Genitiv durch *-e* ausgedrückt wird, was im Falle einer hohen Vorkommenshäufigkeit von femininen Nomen einen höheren Vokalanteil nach sich ziehen würde. D.h. es sind durchaus, bedingt durch die Aufnahme der zu zählenden Einheiten, Unterschiede in der Häufigkeit bestimmter Grapheme „erklärbar".

Dem gegenüber impliziert die Analyse eines Textes, d.h. einer zusammenhängenden natürlich-sprachlichen Äußerung, die zudem als in einer Sprache morphologisch-syntaktisch korrekt gilt, hinsichtlich der Häufigkeit von Graphemen ein etwas anderes Bild. Begründbar wären die Unterschiede u.a. durch die Frequenz von Synsemantika in Texten. Bekanntlich ist in Texten eine hohe Frequenz von Synsemantika zu beobachten, die durch die Notwendigkeit einer morphosyntaktischen Organisation eines Textes bedingt ist. So würde z.B. – um weiterhin beim Slowenischen zu bleiben – in Texten eine hohe Häufigkeit von unterschiedlichen Formen des Auxiliars *biti* (‚sein') wie z.B. *sem, si, je, smo, sta* usw. zu beobachten sein, was einen Einfluss auf die Verteilung von <s>, <e>, <j>, <m>, <t> usw. haben sollte. Des Weiteren kommen in zusammenhängenden Texten Präpositionen wie *in* (‚und'), *v* (‚in'), *za* (‚für') usw. sehr häufig vor, was wiederum einen Einfluss auf die Verteilung von Graphemen in einem Text hat.

---

[2] In den Wörterbüchern einiger Sprachen werden z.B. nur die Grundwörter angegeben und die affigierten Formen sind Teil des Lexems (z.B. Indonesisch); in anderen werden z.B. die meisten Verben in dritter Person, d.h. kürzester Form (nicht im Infinitiv) angegeben, z.B. im Ungarischen. Daher sind wir uns sehr wohl der Tatsache bewusst, dass die von uns gewählte Lemma-Analyse bis zu einem gewissen Grad willkürlich ist, aber z.B. für die Analysen slawischer Sprachen (abgesehen vom Makedonischen bzw. Bulgarischen, wo die Verhältnisse wiederum andere sind) und darauffolgende Vergleiche ebenfalls möglich wären. Bei der Analyse anderer Sprachen und insbesondere bei Vergleichen zwischen diesen Sprachen wäre dies nicht ohne weiteres möglich.

Diese ersten, zum Teil einfach wirkenden, aber aus unserer Sicht durchaus plausiblen Gründe für das unterschiedliche „Verhalten“ von Graphemhäufigkeiten im Wörterbuch bzw. im „Text“ sollen nunmehr etwas ausführlicher und systematischer anhand des Slowenischen überprüft werden. Es geht dabei in diesem Schritt vor allem um die Frage, ob die von uns angeführten Vermutungen bzw. Überlegungen sich in irgendeiner Weise auch empirisch nachweisen lassen oder nicht.

3. Materialbasis: Vorstellung

Für die vorliegende Analyse wurde ein Grund- und Aufbauwortschatz Slowenisch – Deutsch (vgl. Kelih/Vučajnk 2018) herangezogen. Für die im Folgenden durchgeführten Analysen ist nur der slowenische Teil von Interesse. Dieser besteht aus 4950 Lemmata und 5095 begleitenden Beispielsätzen, die insgesamt aus 29764 (orthographischen) Wörtern bestehen. Als Beispiel für den Aufbau des genannten Wörterbuches ist das Lemma *imenovati se* (‚heißen‘, ‚sich nennen‘) anzuführen.

Lemma	imenováti se -újem se *impf*
Beispielsatz	Kakó se imenúje vàš sodélavec?
Übersetzung	Wie heißt Ihr Mitarbeiter?

Die Verbform ist, wie in der slowenischen Lexikographie üblich, als Infinitivform angeführt, inkl. der Spezifizierung des Aspektes (unvollendet) und der Endung für die 1. Person Singular. Der dazugehörige Beispielsatz beinhaltet einen möglichst typischen Verwendungskontext, der dem Lernenden eine leichtere Einprägsamkeit des Lemmas gewährleisten soll.

Für die Bestimmung der Graphemhäufigkeiten[3] sind folgende grundlegende Spezifizierungen von Bedeutung. Unter Graphemen werden die 25 Buchstaben des slowenischen Alphabets (vgl. dazu Kelih 2008, allgemeine Informationen zum slowenischen Graphem- und Phonemsystem vgl. Rehder 2006; Herrity 2010) verstanden. Diese umfassen folgende Elemente: < 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 >. Zwischen Groß- und Kleinbuch-

[3] Eine weitere Besonderheit des untersuchten Wörterbuches ist, dass sowohl die Lemmata als auch die Beispielsätze durchgehend akzentuiert sind. Dies ermöglicht es in Zukunft im untersuchten Material, sowohl die Häufigkeit der Grapheme als auch die der Phoneme inkl. suprasegmentaler Eigenschaften zu bestimmen und zu untersuchen. Für die vorliegende Untersuchung wurden die Betonungen aber nicht weiter berücksichtigt. Sie sind in der geltenden Orthographie nicht üblich und werden nur verwendet, um allfällige Ambiguitäten zu markieren.

staben wird bei der Bestimmung der Graphemhäufigkeiten nicht unterschieden. In den Texten werden allfällig vorkommende „fremde“ Grapheme (wie z.B. <x, y, w> etc.) nicht gezählt. Dies gilt auch für die Interpunktionszeichen.

Hinsichtlich der weiteren Operationalisierungen ist zu erwähnen, dass bei den Graphemanalysen jegliche Annotationen (z.B. Auszeichnung der Wortarten als *f* (feminin), *m* (maskulin), *n* (neutrum), *impf* (unvollendeter Aspekt), *pf* (vollendeter Aspekt), *adj* (Adjektiv) usw.) nicht berücksichtigt werden, sodass nur das zielsprachliche slowenische Material als Lemma bzw. abgeschlossener Beispielsatz analysiert wird. Bei den flektierbaren Wortarten (Verben, Nomina, Adjektive) wird aber neben dem eigentlichen Lemma auch z.B. der Genitiv Singular bei Nomen, die erste Person Singular bei Verben, bei Adjektiven alle drei Genera analysiert. Wie in diesem Beispiel zu sehen ist, würde dies alle kursiv gesetzten Teile betreffen, wie die folgenden Beispiele *imenováti se -újem se* (‚heißen‘), *Japónec -nca* (‚Japaner‘), *lep -a -o* (‚schön‘) zeigen. Die Beispielsätze werden, wie bereits gesagt, als Ganzes analysiert. In einigen wenigen Fällen sind bei ausgewählten Lemmata zwei kontexttypische Sätze zu finden, wobei beide in die Auszählungen einfließen.

3.1. Häufigkeit von Graphemen: Deskription und Modelle

Zu beginnen ist mit der Vorstellung der Graphemhäufigkeiten im Wörterbuch, d.h. also auf Lemma-Ebene. In Tabelle 1 finden sich die absoluten, relativen und prozentuellen Häufigkeiten der ausgezählten Grapheme in alphabetischer Reihenfolge. Rechts in der Tabelle sind die entsprechenden Ranghäufigkeiten (= Sortierung nach ihrer Vorkommenshäufigkeit in dem Rang 1 = häufigstes Graphem, Rang 2 = zweithäufigstes usw.) inkl. der relativen Häufigkeit angegeben.

Tabelle 1: Graphemhäufigkeiten: Wörterbuch

Nr.	Einheiten	F	rel.	%	Einheiten	Rank F.	rel. F.
1	a	7478	0,1490	14,90	a	7478	0,1490
2	b	729	0,0145	1,45	e	5859	0,1167
3	c	709	0,0141	1,41	i	4678	0,0932
4	č	841	0,0168	1,68	o	3907	0,0779
5	d	1319	0,0263	2,63	n	2999	0,0598
6	e	5859	0,1167	11,67	t	2877	0,0573
7	f	100	0,0020	0,20	r	2595	0,0517
8	g	764	0,0152	1,52	s	2325	0,0463
9	h	256	0,0051	0,51	m	1849	0,0368
10	i	4678	0,0932	9,32	l	1803	0,0359
11	j	1406	0,0280	2,80	k	1786	0,0356
12	k	1786	0,0356	3,56	p	1689	0,0337
13	l	1803	0,0359	3,59	v	1643	0,0327
14	m	1849	0,0368	3,68	j	1406	0,0280
15	n	2999	0,0598	5,98	d	1319	0,0263
16	o	3907	0,0779	7,79	z	960	0,0191
17	p	1689	0,0337	3,37	č	841	0,0168
18	r	2595	0,0517	5,17	u	785	0,0156
19	s	2325	0,0463	4,63	g	764	0,0152
20	š	470	0,0094	0,94	b	729	0,0145
21	t	2877	0,0573	5,73	c	709	0,0141
22	u	785	0,0156	1,56	š	470	0,0094
23	v	1643	0,0327	3,27	ž	358	0,0071
24	z	960	0,0191	1,91	h	256	0,0051
25	ž	358	0,0071	0,71	f	100	0,0020
		50158	1	100		50158	

Um im Folgenden einen besseren Vergleich der Daten zu gewährleisten, werden diese in Abb. 1 in der Form von relativen Anteilen präsentiert.

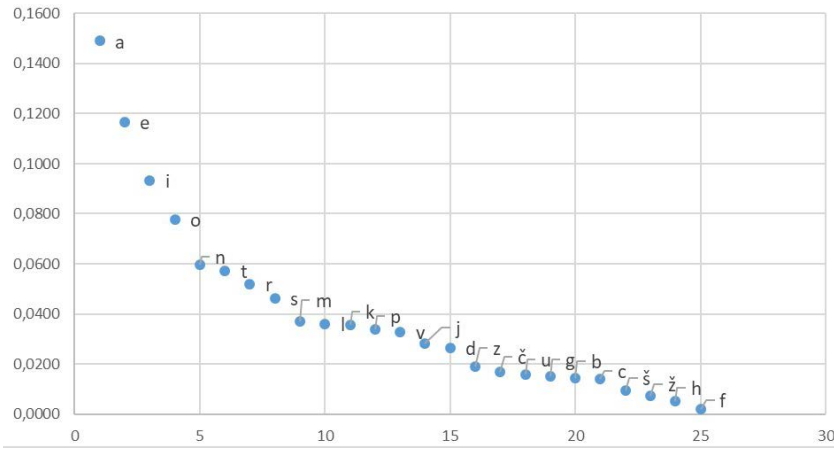

Abb. 1: Graphemhäufigkeiten im Slowenischen: Wörterbuch (Lemmata)

Die Rangverteilung wird eindeutig – und dies ist schön aus Abb. 1 abzulesen – von Vokalen dominiert. Hierbei ist im Wörterbuch <a> das häufigste Graphem (Anteil von 0,149), welches „gefolgt" wird von <e> mit einem Anteil von 0,1167, <i> mit einem Anteil von 0,0932 und <o> mit einem Anteil von 0,0779. Zählt man diese häufigsten Einheiten zusammen, so ergibt sich, dass die vier Grapheme <a, e, i und o> bereits 43% aller Grapheme im Wörterbuch, d.h. der untersuchten 4950 Lemmata ausmachen. Hervorzuheben ist die Frequenz des Vokals <u>, welche im Vergleich zu den anderen Vokalen sehr gering ausfällt und gerade einmal einen relativen Anteil von 0,0156 einnimmt. Dies ändert aber nichts an dem Befund, dass der vordere Rangbereich eindeutig von Vokalen dominiert wird.

Von Interesse sind auch die Häufigkeiten der Konsonanten. Die häufigsten Konsonanten (in Klammern ist nunmehr nur der jeweilige Rang angeführt, die entsprechenden relativen Anteile können Tabelle 1 entnommen werden) sind <n> (5), <t> (6), <r> (7), <s> (8), <m> (9) und <l> (10). Auffällig an dieser Verteilung ist, dass im vorderen Rangbereich vor allem diejenigen dominieren, die die entsprechenden Sonore /n, r, m/ und /l/ repräsentieren. Es ist somit – und dies ist durchaus bemerkenswert – innerhalb der Konsonanten eine eindeutige Präferenz für sonore Laute festzustellen. In Anbetracht der Tatsache, dass sich die Analysen auf Wörterbuchmaterial beziehen und neben den Stammmorphemen inkl. Präfixe vor allem die jeweiligen Endungen analysiert werden, können die bisher genannten 10 Grapheme <a, e, i, o, n, t, r, s, m und l> als zentrale (morphologische) „Systemeinheiten" des Slowenischen bezeichnet werden. In

jedem Fall sind es oft gebrauchte – der relative Anteil macht fast drei Viertel aller Grapheme aus (0,7247) – und somit wichtige Bausteine für die einzelnen Lemmata. Ein Blick auf das Ende der Ranghäufigkeitsverteilung bringt keine Überraschungen mehr zu Tage, wo insgesamt die verbleibenden 15 Einheiten zusammen genommen einen relativen Anteil von 0,2753 ausmachen. Es handelt sich vor allem um Grapheme, die Affrikate, Frikative u.ä. repräsentieren, die offenbar einen eher peripheren Status einnehmen (eventuell wäre dies durch die Komplexität der Artikulation zu erklären). Zu berücksichtigen wäre auch die Sprachgeschichte der slawischen Sprachen, wonach das Graphem /f/ in der Regel kein genuines Phonem slawischer Sprachen repräsentiert, sondern nur in Lehnwörtern vorkommt. Aus dieser Perspektive ist es verständlich, dass das <f> in unserem Sample den letzten Rangplatz einnimmt.

Nachdem in Grundzügen die deskriptiven Eigenschaften der untersuchten Graphemhäufigkeiten skizziert wurden, ist in einem nächsten Schritt die Frage zu klären, ob und in welcher Form diese auf der Modellebene zu erfassen sind. Es ist in der quantitativen Linguistik üblich, die Rangverteilung von linguistischen Entitäten mit Hilfe einer Zipf-Funktion zu erfassen. Etwas bessere Resultate kann man mit der Mandelbrot-Funktion bzw. auch mit der Zipf-Alekseev-Funktion bekommen. Die Rechnung im slowenischen Bereich – das kann für alle Sprachen gelten – hat den Nachteil, dass die letzten Grapheme sehr selten vorkommen und dadurch die Sequenz verzerrt wird. Aus diesem Grund wird die Menzerath-Funktion (vgl. zur Ableitung Kolenčíková/Altmann 2020: 89) gewählt,[4] $y = a*x^{b}*\exp(-c*x)$, die die besten Resultate liefert. Das Resultat der Anpassung ist in Tabelle 2 dargestellt.

Tabelle 2
Rangierte Graphemhäufigkeiten Slowenisch (Wörterbuch)

Rang	Abs. F.	theo. Menzerath F.
1	7478	7545,61
2	5859	5648,83

[4] Damit ist kein Anspruch gestellt, dass dies das einzig adäquate Modell wäre. Es werden viele unterschiedliche Modelle in diesem Zusammenhang diskutiert, wobei nach wie vor mehr oder weniger unklar ist, welche dahinterliegenden generierenden Mechanismen dafür verantwortlich sind. Klar ist, dass Faktoren wie die Inventargröße, das Ausmaß an distributionellen Beschränkungen, die Silbenstruktur bzw. auch die Wortlänge als Einflussgrößen anzusehen sind. Eine Erweiterung der in Frage kommenden Modelle ergibt sich, wenn man auch diskrete Verteilungen in Betracht zieht. Bislang ungeklärt ist auch die Frage der Parameterinterpretation. Im obigen Fall werden sie iterativ aus den empirischen Daten geschätzt.

3	4678	4615,17
4	3907	3908,05
5	2999	3374,81
6	2877	2950,70
7	2595	2602,70
8	2325	2309,06
9	1849	2058,87
10	1803	1842,73
11	1786	1654,33
12	1689	1488,93
13	1643	1342,92
14	1406	1213,41
15	1319	1098,09
16	960	995,08
17	841	902,81
18	785	819,96
19	764	745,41
20	729	678,20
21	709	617,53
22	470	562,67
23	358	513,01
24	256	468,00
25	100	427,17
a=8165,02, b=-0,3039, c=0,0789, R^2=0,99		

Trotz der stellenweise großen Unterschiede zwischen den empirischen Frequenzen und den Werten der Menzerath-Funktion deutet der Determinationskoeffizient auf eine sehr gute Übereinstimmung hin. D.h. man kann davon ausgehen, dass das propagierte Modell ein geeignetes ist, um die slowenischen Graphemhäufigkeiten auf Wörterbuchebene in entsprechender Weise zu modellieren.

In einem nächsten Schritt können nun die entsprechenden Graphemhäufigkeiten auf der Textebene im Detail besprochen werden. Dabei soll analog vorgegangen werden, wie im Fall der oben präsentieren Häufigkeiten auf Wörterbuchebene. Zu beginnen ist mit den entsprechenden Rohdaten, die in Tabelle 3

zu finden sind. Zuerst sind die absoluten, relativen, prozentuellen Häufigkeiten in alphabetischer Reihenfolge angegeben, dann die entsprechenden Ranghäufigkeitsverteilungen.

Tabelle 3: Graphemhäufigkeiten: Slowenische Texte (Beispielsätze)

Nr.	Einheiten	F	rel.	%	Einheiten	Rang F.	rel. F.
1	a	15341	0,1030	10,30	e	15890	0,1067
2	b	2650	0,0178	1,78	a	15341	0,1030
3	c	1363	0,0092	0,92	o	14298	0,0960
4	č	2177	0,0146	1,46	i	12798	0,0859
5	d	4872	0,0327	3,27	n	9028	0,0606
6	e	15890	0,1067	10,67	l	7784	0,0523
7	f	268	0,0018	0,18	s	7716	0,0518
8	g	1801	0,0121	1,21	r	7683	0,0516
9	h	1190	0,0080	0,80	j	6874	0,0462
10	i	12798	0,0859	8,59	t	6511	0,0437
11	j	6874	0,0462	4,62	v	6103	0,0410
12	k	5251	0,0353	3,53	p	5426	0,0364
13	l	7784	0,0523	5,23	k	5251	0,0353
14	m	4721	0,0317	3,17	d	4872	0,0327
15	n	9028	0,0606	6,06	m	4721	0,0317
16	o	14298	0,0960	9,60	z	3384	0,0227
17	p	5426	0,0364	3,64	u	2958	0,0199
18	r	7683	0,0516	5,16	b	2650	0,0178
19	s	7716	0,0518	5,18	č	2177	0,0146
20	š	1743	0,0117	1,17	g	1801	0,0121
21	t	6511	0,0437	4,37	š	1743	0,0117
22	u	2958	0,0199	1,99	c	1363	0,0092
23	v	6103	0,0410	4,10	h	1190	0,0080
24	z	3384	0,0227	2,27	ž	1073	0,0072
25	ž	1073	0,0072	0,72	f	268	0,0018
		148903	1	100		148903	1

Erste Auffälligkeit ist (vgl. dazu Abb. 2), dass nunmehr <e> auf Textebene als das häufigste Graphem erscheint. Der nächste Befund ist, dass der relative

Anteil des häufigsten Graphems durchaus geringer (nunmehr 0,1030) ist als im Fall der Wörterbuchdaten. Betrachtet man die ersten vier Rangplätze, so ist aber zu bemerken, dass die häufigsten Grapheme wiederum alle Vokale repräsentieren. Es gibt zwar einige Verschiebungen hinsichtlich des Rangplatzes, aber tendenziell bleibt das Gesamtbild erhalten, welches bereits aus der Analyse der Wörterbuchdaten bekannt ist: vier Vokale <e, a, o und i> dominieren eindeutig den vorderen Ranghäufigkeitsbereich und nehmen einen relativen Gesamtanteil von ca. 0,40 ein. Im Vergleich zu den Wörterbuchdaten ist dies ein etwas geringerer Anteil. Somit gilt festzuhalten, dass durchaus kleinere Verschiebungen zwischen der Wörterbuch- und Textanalyse zu beobachten sind. Für eine graphische Darstellung der relativen Ranghäufigkeiten in den Beispielsätzen siehe Abb. 2.

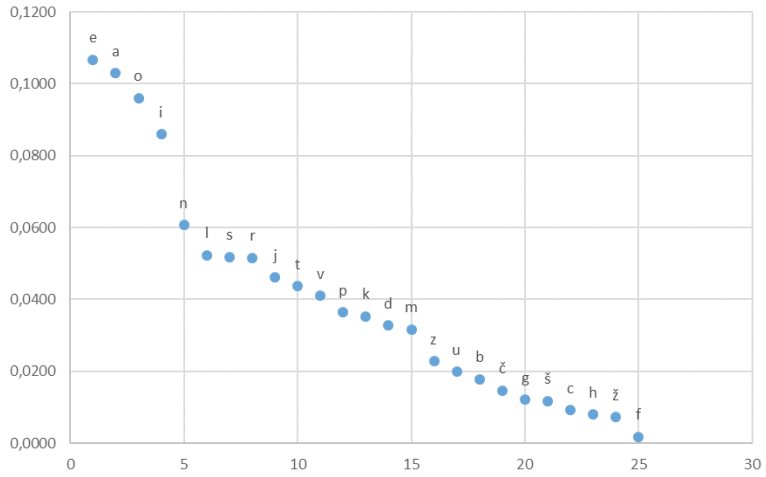

Abb. 2: Graphemhäufigkeiten im Slowenischen: Textebene

Darüber hinaus ergeben sich aber vor allem bei den Konsonanten einige auffällige Verschiebungen der Rangreihenfolge. Nunmehr treten als die sechs häufigsten Konsonanten <n, l, s, r, j> und <t> in Erscheinung, die alle in etwa je einen gleich hohen relativen Anteil von 0,05 haben. Eine markante Verschiebung im Vergleich zu den Wörterbuchdaten ergibt sich auch bei der relativ hohen Vorkommenshäufigkeit von <j>. Im Falle der Wörterbuchdaten hat <j> einen Anteil von 0,028, während sich nun im Fall der Textdaten der Anteil auf 0,046 fast verdoppelt. Dieses Phänomen kann, wie bereits einleitend angedeutet, dadurch erklärt werden, dass in den Textdaten eine hohe Frequenz von Synsematika zu verzeichnen ist, und dort <j> überdurchschnittlich oft (z.B. in

der Form des Auxiliars *biti* („sein“) in der 3. Person Sg. *je*, bei Pronomen *jo*, *svoj* usw.) vorkommt. Eine mögliche Erklärung, die ohne Zweifel in Zukunft noch genauer zu überprüfen sein wird.

Betrachtet man abschließend den relativen Anteil von den 10 häufigsten Graphemen, so ergibt sich ein Anteil von 0,6979, was im Vergleich zu den Wörterbuch-Daten (0,7247) um etwas geringer ist. Dies kann dahingehend interpretiert werden, dass es auf der Textebene vermutlich zu einer geringfügig höheren Gleichverteilung der untersuchten Einheiten kommt. Dieser Befund wird aber im Kapitel 2.3. noch näher zu besprechen sein.

In einem nächsten Schritt ist nun die Frage zu klären, ob auch für die Textdaten die oben verwendete Menzerath-Funktion herangezogen werden kann oder nicht. Die Resultate der Modellierung finden sich in Tabelle 4.

Tabelle 4
Slowenische Graphemhäufigkeiten: Textebene

Rang	Frequenz	Menzerath f.
1	15890	16422,45
2	15341	14842,92
3	14298	13386,84
4	12798	12063,11
5	9028	10865,09
6	7784	9783,10
7	7716	8807,03
8	7683	7927,13
9	6874	7134,32
10	6511	6420,02
11	6103	5777,14
12	2426	5198,17
13	5251	4676,98
14	4872	4207,87
15	4721	3785,66
16	3384	3405,71
17	2958	3063,80
18	2650	2756,14
19	2177	5479,33
20	1801	2230,27

21	1743	2006,19
22	1363	1804,59
23	1190	1623,23
24	1073	1460,08
25	268	1313,30
a=18263,30, b=-0,00738, c=0,1062, R^2=0,9703		

In dieser Form der Analyse ist zwar die Diskrepanz zwischen empirischen und theoretischen Häufigkeiten vor allem in den unteren Rangplätzen durchaus groß, aber der Determinationskoeffizient (R^2=0.97) deutet weiterhin auf eine akzeptable Anpassung hin. Das geringfügig schlechtere Anpassungsergebnis im Vergleich zu den Wörterbuchdaten lässt sich eventuell dadurch erklären, dass es sich hierbei um keine „abgeschlossenen" Texte handelt, sondern um Beispielsätze, deren Struktur durch das Vorkommen des vorgegebenen Lemmas gewissermaßen „prädestiniert" ist.

Die bisher präsentierten Daten mögen auf der Modellebene auf den ersten Blick keine gravierenden Unterschiede zwischen den Graphemhäufigkeiten auf Wörterbuch- und Textebene ergeben haben. Eine nähere Betrachtung ist aber durchaus lohnenswert. In erster Linie ist der unterschiedliche Ausnutzungsgrad der einzelnen Elemente, wie z.B. des jeweils häufigsten Graphems, im Wörterbuch im Vergleich zum Textmaterial von Bedeutung. So hat z.B. im Wörterbuch das Graphem <a> einen Anteil von 0,149, während in den Textdaten das häufigste Graphem <e> einen insgesamt geringeren Anteil von 0,1167 hat. Aus mathematischer Sicht ist es an dieser Stelle notwendig, diese bemerkbaren Differenzen auf statistische Signifikanz hin zu überprüfen. Üblicherweise würde man bei derartigen Fragestellungen auf den χ^2 -Test Bezug nehmen. Es ist aber zu bemerken, dass die Häufigkeiten zu groß sind und somit im Grunde einen „sinnlosen", d.h. nicht ordnungsgemäß interpretierbaren χ^2-Wert ergeben würden. Daher pflegt man mit solchen großen Daten wie den vorliegenden als sinnvolle Alternative einen parameterfreien Rangvergleichstest durchzuführen, indem man die Ränge der gleichen Grapheme vergleicht und auswertet.

Wir führen daher einen Rangkorrelationstest durch, der in jedem Lehrbuch der Statistik (z.B. Zöfel 2012) enthalten und beschrieben ist. Wie in Tabelle 5 zu sehen ist, wurden jedem Graphem der beiden Stichproben (Text und Wörterbuch) die Rangnummern zugeschrieben.

Tabelle 5
Rangkorrelationstest für Grapheme in Text und Wörterbuch

Graphem	Rang Wörterbuch	Rang im Text	d	d2
a	1	2	-1	1
b	20	18	2	4
c	21	22	-1	1
č	17	19	-2	4
d	15	14	1	1
e	2	1	1	1
f	25	25	0	0
g	19	20	-1	1
h	24	23	1	1
i	3	4	-1	1
j	14	9	5	25
k	11	13	-2	4
l	10	6	4	16
m	9	15	-6	36
n	5	5	0	0
o	4	3	1	1
p	12	12	0	0
r	7	8	-1	1
s	8	7	1	1
š	22	21	1	1
t	6	10	-4	16
u	18	17	1	1
v	13	11	2	4
z	16	16	0	0
ž	23	24	-1	1
				122

Setzt man die erhaltenen Zahlen in die Formeln ein, wobei n = 25, $\sum_{i=0}^{n} d_i^2 = 122$, wie aus der Tabelle 5 ersichtlich, dann erhält man

$$r = 1 - \left(\frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}\right) = 1 - 6(122)/[25*(252-1)] = 0.9530$$

Das Resultat testet man mit einem t-Test mit n-2 = 23 Freiheitsgraden, indem man r in die Formel

$$t = \frac{r}{\sqrt{(1-r^2)(n-2)}}$$

einsetzt.

Wir erhalten $t = 0.9530/\sqrt{(1 - 0.953^2)(25 - 2)} = 0.8831$ woraus folgt, dass der Unterschied zwischen den Rängen der Graphemhäufigkeiten im Wörterbuch und in den Texten nicht signifikant ist. Trotz der zweifellos vorhandenen quantitativen Unterschiede zwischen den Daten aus dem Wörterbuch und den Texten sind diese im statistischen Sinn – sofern man den oben angeführten Test verwendet – statistisch nicht signifikant. Somit lässt sich folgendes wichtige Zwischenergebnis festhalten: Die Dichotomie von Wörterbuch und Text, die für viele quantitative linguistische Untersuchungen als geradezu konstituierend angesehen wird, lässt sich zumindest in dem hier untersuchten slowenischen Material nicht „nachweisen". Es zeigen sich, zumindest auf der deskriptiven Ebene, zwar bestimmte Unterschiede in Hinblick auf die Rangposition bestimmter Grapheme. Allerdings sind diese Unterschiede quantitativ nicht so groß, als dass damit auch statistisch relevante Differenzen einhergehen würden. Dieses durchaus ähnliche Rangierungsverhalten, sowohl im Wörterbuch als auch im Text, zeigt sich zudem auch daran, dass für beide Arten sich jeweils das gleiche theoretische Modell als geeignet erwiesen hat.

Selbstverständlich gilt aber dieser Befund, der auf eine Irrelevanz der Unterscheidung von Wörterbuch vs. Text hinweisen würde, nur für die hier verwendeten Daten. Diese sind, und dies muss schon gesondert betont werden, dadurch gekennzeichnet, dass die angeführten Lemmata auch in den jeweiligen Beispielsätzen vorkommen. In Bezug auf das quantitative Ausmaß der Übereinstimmung von Wörterbuch und Text lässt sich sagen, dass klarerweise die 4950 Lemmata ein Teil der 29764 Wörter der Beispielsätze sind, was einem Anteil von ca. 16% entspricht. Darüber hinaus ist es nicht ausgeschlossen, dass viele der Lemmata, die zum Grund- und Aufbauwortschatz gehören, auch in entsprechend hoher Frequenz in den Beispielsätzen vorkommen. Insofern ist das Ergebnis des Signifikanztestes tatsächlich als plausibel anzusehen.

Darüber hinaus kann aber – wie nun im folgendem Kapitel zu zeigen sein wird – die Verteilung von Graphemen im Wörterbuch und im Text in einem etwas anderen Licht gesehen werden, indem man nicht nur die Häufigkeit der einzelnen Grapheme, sondern die Häufigkeit von Vokalen und Konsonanten in Betracht zieht.

3.2. Vokal- und Konsonantenhäufigkeiten – Unterschiede?

Eine weitere Möglichkeit des Vergleiches der Graphemhäufigkeiten auf Wörterbuch- und Textebene besteht darin, die untersuchten Einheiten auf zwei Klassen, nämlich Vokale und Konsonanten, zu reduzieren. Dies sollte einen „binären" Blick auf den unterschiedlichen Ausnutzungsgrad von Vokalen und Konsonanten zulassen. Als vokalische Einheiten werden im Slowenischen <i, e, a, o, u> gezählt[5] und auf dieser Basis wird die Häufigkeit bzw. der relative Anteil von Vokalen im Wörterbuch und in den Beispielsätzen bestimmt; alle nicht-vokalischen Elemente ergeben (1-V) die Gruppe der Konsonanten. Damit erhält man eine wichtige Information über die funktionale Auslastung dieser beiden Klassen.

Im Falle des Slowenischen als einer indoeuropäischen Sprache aus der Familie der slawischen Sprachen lässt sich feststellen, dass auf Lemma-Ebene (vgl. dazu Tabelle 6 mit den entsprechenden Rohdaten) über 45% der Vokale zu finden sind. Im Vergleich dazu beträgt der Anteil von Vokalen in den Beispielsätzen nur etwa ca. 41%. Somit hat man es auf den ersten Blick in der Tat mit einem recht unterschiedlichen Ausnutzungsgrad von Vokalen auf Wörterbuch- bzw. Textebene zu tun.

Dies lässt den vorläufigen Schluss zu, dass offenbar die morphologische Information im Slowenischen stärker von Vokalen getragen wird. Dies lässt sich auch leicht aus dem analysierten Material der Lemmata ableiten (man denke z.B. an den Genitiv von Nomina bei Feminina, der vornehmlich durch <-e> ausgedrückt wird, während er bei männlichen Nomina im Genitiv <-a> ist). Der unterschiedliche Anteil von Vokalen und Konsonanten und die entsprechenden Häufigkeiten lassen sich aus der untenstehenden Tabelle 6 gut ablesen.

---

[5] Auf eine gesonderte Zählung von <r> welches in bestimmten Positionen ebenfalls als silbenbildend aufgefasst werden kann, wird im gegebenen Zusammenhang verzichtet. Es kommt dieses in dieser Funktion auch nur in einer äußerst geringen Häufigkeit vor. Es ist zu bemerken, dass selbst eine Trennung von Vokalen und Konsonanten nicht in jeder Sprache eindeutig möglich ist. Man denke an Halbvokale, Diphthonge und ähnliches.

Tabelle 6
Vokal- und Konsonantenanteil (Wörterbuch, Text)

	Wörterbuch		Text	
	abs.	%	abs.	%
Vokale	22707	0,4526	61285	0,4129
Konsonanten	27478	0,5475	87618	0,5884
Summe	50185	1	148903	1

Aber auch in diesem Fall kann ein entsprechender statistischer Test genauere Informationen über signifikante Unterschiede geben. Vergleicht man die Anteile von Vokalen in Wörterbuch und Text, so benutzt man einen sogenannten Normaltest: man berechnet $h = (x_1 + x_2)/(n_1 + n_2) = (22707 + 61285)/(50185 + 148903) = 0.4219$. Diese Werte setzt man in die Formel

$$u = \frac{p_{1-p_2}}{\sqrt{h(1-h)(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{0.4526-0.4129)}{\sqrt{0.4219(1-0.4219)*(\frac{1}{50185}+\frac{1}{148903})}} = 15.57,$$

was einen hoch signifikanten Unterschied anzeigt. Damit wäre abgesichert, dass der Unterschied des Vokalanteiles zwischen dem „Wörterbuch“ und dem „Text“ entsprechend statistisch signifikant ist.

Ob diese tendenziell unterschiedliche Verteilung von Vokalhäufigkeiten auch für andere Sprachen gilt, muss natürlich im Detail erst untersucht werden. Verwiesen sei auf ältere Überlegungen von Skalička (1966: 114), der bereits einen Zusammenhang zwischen dem Vokalanteil und der Funktion der einzelnen Vokale anspricht. Vokale haben nicht nur eine akustische und silbenbildende Funktion, sondern treten seiner Meinung nach eher als Träger von morphologischer Information auf Ebene des Wörterbuchs in Erscheinung. In jedem Fall sind aber noch weitere Untersuchungen innerhalb einer Sprache notwendig, indem man z.B. den Vokalanteil separat in Nomina, Verben, Adjektiven usw. vergleichen würde, und erst dann würde man sehen, ob und welche Unterschiede damit einhergehen. Dies könnte dazu beitragen, die morphologisch-grammatische Dimension von Vokalen bzw. Konsonanten besser einzuschätzen helfen. Der unterschiedliche Ausnutzungsgrad der Grapheme wird in Kap. 3.3 mit Hilfe der Wiederholungsrate thematisiert.

3.3. Vergleich der Wiederholungsrate: Wörterbuch vs. Text

Ein wichtiges Merkmal sprachlicher Systeme im Allgemeinen ist es, dass deren Einheiten (Phoneme, Grapheme, Silben, Morpheme usw.) in Texten nicht mit der gleichen Häufigkeit (= Gleichverteilung) vorkommen, sondern bestimmte

Einheiten überdurchschnittlich oft genutzt werden. Die linguistische Operationalisierbarkeit und verfeinerte Metrisierung der funktionalen Belastung ergibt sich in Bezug auf die Häufigkeit von Phonemen durch die Berechnung der sogenannten Wiederholungsrate. Die Wiederholungsrate (*RR*) ist eine in der quantitativen Phonologie häufig diskutierte Kenngröße (vgl. Altmann/Lehfeldt 1980: 151f). Diese lässt sich als

$$RR = \sum_{r=1}^{n} p_r^2$$

d.h. als die Summe der quadrierten relativen Häufigkeiten (*p*) der einzelnen Phoneme (bzw. Grapheme) berechnen. Die Wiederholungsrate ist ein Maß der Gleichverteilung der Häufigkeiten von Phonemen und besagt etwas über den Ausnutzungsgrad der einzelnen Phoneme. Grundlegendes Merkmal der Wiederholungsrate ist es, dass sie umso kleiner wird, je ähnlicher (gleicher) die Häufigkeiten von Phonemen verteilt sind. Ähnlich bedeutet in diesem Zusammenhang mit einer gleichen bzw. annähernd gleichen Häufigkeit vorkommend.

In Bezug auf die hier verfolgte Fragestellung ist davon auszugehen, dass auf der Textebene die Wiederholungsrate im Vergleich zum Wörterbuch geringer sein sollte. Dies lässt sich dadurch begründen, dass es im Wörterbuch und aufgrund der standardisierten Darstellung der Lemmata zu einer überdurchschnittlichen Auslastung ausgewählter Einheiten kommt. Dies ergibt sich vor allem daraus, dass für die morphologische Kodierung auf Lemma-Ebene sehr oft die gleichen Grapheme in Frage kommen, während auf der Textebene durch das Einführen von flektierenden Formen die Bandbreite von in Frage kommenden Graphemen erhöht wird. Wie aus der untenstehenden Tabelle 7 zu entnehmen ist, lässt sich tatsächlich zeigen, dass die *RR* auf Textebene etwas niedriger ausfällt.

Tabelle 7: Wiederholungsrate für Wörterbuch vs. Text

	Lemma	Beispielsätze
RR	0,0716	0,0626

Die Vorhersage hinsichtlich der Wiederholungsrate lässt sich somit bestätigen. Die Wiederholungsrate ist bei der jeweiligen Wörterbuchanalyse höher als bei den Häufigkeiten auf der Basis der Textbeispiele.

Auch in diesem Fall kann ein entsprechender Signifikanztest durchgeführt werden, der Auskunft über tatsächlich relevante Unterschiede gibt. Für die Untersuchung der Unterschiede der Wiederholungsrate und der Varianz in den beiden Stichproben kann ein asymptotischer Test verwendet werden, der in Popescu et al. (2009: 165-169) detailliert beschrieben ist. Unter Anwendung dieser Methode ergibt sich ein z-Wert = 20,71, was im Fall eines Signifikanz-

niveaus von $\alpha = 5\%$ und einem klar über $z > 1{,}96$ liegenden Wert auf deutliche Unterschiede zwischen der Wiederholungsrate im Wörterbuch und in den Textbeispielen hindeutet. Somit lässt sich zeigen, dass nicht nur der Vokal- und Konsonantenanteil statisch signifikant unterschiedlich ausfällt, sondern dass das auch auf der Ebene der Wiederholungsrate zum Ausdruck kommt.

4. Zusammenfassung

In der vorliegenden Untersuchung wird der Frage nachgegangen, inwiefern bei quantitativen Graphemuntersuchungen der Unterschied zwischen Daten aus Wörterbüchern bzw. aus Texten von Relevanz sein kann. Ohne Zweifel ist diese Dichotomie für die Linguistik von nachhaltiger Bedeutung und ist als grundlegende Unterscheidung nicht in Zweifel zu ziehen. Betrachtet man allerdings die empirische Dimension dieser Fragestellung in Zusammenhang mit der Frage von Graphemhäufigkeiten, so ergeben sich auf der Basis der hier untersuchten slowenischen Daten folgende vorläufige Resultate bzw. Befunde:

1. Sowohl im Wörterbuch als auch in Texten treten die Vokale als die häufigsten Einheiten auf, wenngleich diese qualitativ gesehen unterschiedlich sein können. Im Vergleich ergeben sich jeweils leichte Verschiebungen in der Häufigkeit hinsichtlich einzelner Vokale bzw. Konsonanten.
2. Die Ranghäufigkeiten von beiden Stichproben lassen sich mit ein und demselben theoretischen Modell erfassen.
3. Aus statistischer Sicht lässt sich auf der Basis der absoluten Ranghäufigkeiten kein Unterschied zwischen Daten aus dem Wörterbuch und den Texten nachweisen. Dies ist in unserem Fall sicherlich der spezifischen Datenauswahl geschuldet.
4. Im Gegensatz zu den Befunden, die die einzelnen Ranghäufigkeiten betreffen, lässt sich zeigen, dass (a) der Vokal- bzw. Konsonantenanteil und (b) die Wiederholungsrate im Wörterbuch sich statistisch signifikant von der im Text unterscheidet.

Zusammenfassend lässt sich somit im Grunde festhalten, dass sich kein eindeutiges Bild gewinnen lässt. Einerseits lassen sich starke Hinweise finden, dass die Unterscheidung zwischen Wörterbuch und Text insbesondere in Hinblick auf den Vokal- bzw. Konsonantenanteil von Bedeutung sein kann, während die theoretische Rangverteilung auf eine in etwa ähnliche Verteilung zu deuten scheint. Selbstverständlich können diese Befunde vor allem dem hier verwendeten Material geschuldet sein, zumal es, wie mehrfach betont, eine enge Verbindung zwischen den untersuchten Lemmata und den Beispielsätzen zu beachten gilt. Abgesehen davon, dass andere Formen der Operationalisie-

rung von *Wörterbuch* bzw. *Text* vorstellbar sind, ergeben sich eine Reihe von notwendigen Folgeuntersuchungen. Aus unserer Sicht ist es eine zentrale Perspektive die Häufigkeit von Graphemen in einzelnen Wortarten und auch Textsorten, aber auch differenziertes Wörterbuchmaterial zu untersuchen. Nicht vernachlässigt werden sollte auch die Frage der Stichprobengröße, welches ein zentrales und bis heute nur in Ansätzen gelöstes Problem der Zählung und Analyse von Graphemen und Phonemen ist. Darüber hinaus – und auf das hatte N.S. Trubeckoj ebenfalls deutlich verwiesen – wird der Kombinatorik von Graphemik mehr Aufmerksamkeit zu schenken sein.

Literatur

Altmann, G. (1980): „Prolegomena to Menzerath's law". In: Grotjahn (Hrsg.) (1980), 1-10.

Altmann, G.; Fan, F. (Hrsg.) (2008): *Analyses of Script. Properties of Characters and Writing Systems.* Berlin, New York. (= Quantitative Linguistics, 63)

Altmann, G.; Lehfeldt, W. (1980): *Einführung in die quantitative Phonologie*. Bochum. (= Quantitative Linguistics, 7)

Coloma, G. (2015): „The Menzerath-Altmann Law in a Cross-Linguistic Context", in: *Sky Journal of Linguistics* (28), 139-159.

Grotjahn, R. (Hrsg.) (1980): Glottometrika 2. Bochum. (= Quantitative Linguistics, 3)

Grzybek, P.; Kelih, E. (2005): „Häufigkeiten von Buchstaben/Graphemen/Phonemen: Konvergenzen des Rangierungsverhaltens", in: *Glottometrics* (9), 62-74.

Grzybek, P.; Kelih, E.; Stadlober, E. (2009): „Slavic Letter Frequencies: A Common Discrete Model and Regular Parameter Behavior?" In: Köhler (Hrsg.) (2009), 17-33.

Herrity, P. (2010): *Slovene. A comprehensive grammar*. London.

Kelih, E. (2012): „Systematic interrelations between grapheme frequencies and word length: Empirical evidence from Slovene", in: *Journal of Quantitative Linguistics* 19 (3), 205-231.

Kelih, E. (2008): „The phoneme-grapheme relationship in Slovene". In: Altmann et al. (Hrsg.) (2008), 61-74.

Kelih, E.; Vučajnk, T. (2018): *Slovensko-nemški tematski slovar: osnovno in razširjeno besedišče 4500 gesel, frazemov in stavčnih primerov. Grund- und Aufbauwortschatz Slowenisch-Deutsch. 4500 Lemmata, Phrasen und Satzbeispiele*. Klagenfurt/Celovec, Wien/Dunaj.

Köhler, R. (2005): „Synergetic linguistics“. In: Köhler et al. (Hrsg.) (2005), 760-774.

Köhler, R. (Hrsg.) (2009): *Issues in Quantitative Linguistics*. Lüdenscheid. (= Studies in Quantitative Linguistics, 5)

Köhler, R.; Altmann, G.; Piotrowski, R.G. (Hrsg.) (2005): *Quantitative Linguistics. An International Handbook.* New York, Berlin.

Kolenčíková, N.; Altmann, G. (2020): „Analysis of Prepositions in Marína (Slovak Romantic Poem)“, in: *Glottometrics* (48), 88-107.

Martindale, C.; Gusein-Zade, S.; McKenzie, D.P; Borodovsky, M.Y (1996): „Comparison of equations describing the ranked frequency distributions of graphemes and phonemes“, in: *Journal of Quantitative Linguistics* (3), 106-112.

Rehder, P. (2006): „Das Slovenische“. In: Peter Rehder (Hrsg.): *Einführung in die slavischen Sprachen (mit einer Einführung in die Balkanphilologie)*. 5. Auflage. Darmstadt, 230-245.

Skalička, V.(1966): „Konsonatenkombinationen und linguistische Typologie“, in: *Travaux Linguistiques de Prague* (1), 111-114.

Trubetzkoy, N.S. (1989) [1939]: *Grundzüge der Phonologie.* 7. Auflage. Göttingen. (= Travaux du Cercle Linguistique de Prague, 7)

Zöfel, Peter (2002): *Statistik verstehen. Ein Begleitbuch zur computergestützten Anwendung*. München u.a.

emmerich.kelih@univie.ac.at