

Das Grazer Projekt zu Wortlängen(häufigkeiten)

Emmerich Kelih, Peter Grzybek, Ernst Stadlober¹

1. Projektbeschreibung: Allgemeiner Hintergrund

Mit 1.4.2002 fördert der Österreichische *Fonds zur Förderung der wissenschaftlichen Forschung* (FWF, Wien; Projektnummer: P-15485) finanziell ein auf drei Jahre ausgerichtetes Projekt zur Erforschung von Wortlängen(häufigkeiten) in Texten slawischer Sprachen. An dem interdisziplinären und interuniversitären Projekt, das verantwortlich von Peter Grzybek (Institut für Slawistik, Universität Graz) in Zusammenarbeit mit Ernst Stadlober (Institut für Statistik, TU Graz) geleitet wird, arbeiten SpezialistInnen aus Textwissenschaft, Informatik und Statistik mit.

Der allgemeine theoretische Hintergrund des Projekts ist ziemlich weitreichend und anspruchsvoll: Man geht von nicht mehr und nicht weniger aus, als dass die komplementäre Anwendung quantitativer und qualitativer Methoden im Bereich der Geistes- bzw. Kulturwissenschaften eine Möglichkeit darstellt, den immer noch lebendigen Mythos von den "zwei Kulturen" zu überwinden, wie er in den 50er Jahren von Snow ins Leben gerufen und von seinen Anhängern immer wieder künstlich wiederbelebt wurde. Sobald man jedoch 'Natur' und 'Kultur' als spezifische (kulturelle) Konstrukte versteht, ändert sich die Sichtweise und es zeichnen sich zwei Perspektiven ab, wie man mit der vermeintlichen Gegenüberstellung dieser beiden Konzepte umgehen könnte: Einerseits lassen sich die historisch wechselnden Definitionen von 'Kultur' und 'Natur' (und diskursstrategische Gründe für diese Definitionen) selbst zu einem wissenschaftlichen Thema machen, andererseits muss es dann darum gehen, die Konvergenzen zwischen 'Natur' und 'Kultur' selbst zu fokussieren. Und genau diese Annahme kennzeichnet den theoretischen Hintergrund des Grazer Projekts: Während Sprache und sprachlicher Text als spezifische kulturelle Produkte und als spezifische Zeichensysteme innerhalb kultureller Gefüge angesehen werden, lassen sich die zu ihrer Analyse angewendeten statistischen Verfahren als geeignete Meta-Sprache für Kulturstudien im allgemeinen, für linguistische Studien als einer von deren Spezialbereichen im besonderen ansehen.

In Anbetracht dieser eher verwegenen, wenn nicht vermessenem allgemeinen Annahmen nimmt sich das konkrete Forschungsziel des Grazer Projekts vergleichsweise bescheiden aus, insofern "nur" Wortlängen, Wortlängenhäufigkeiten und Faktoren, die hierauf Einfluss haben, betrachtet werden. Ungeachtet dessen ist die eingeschlagene Perspektive innovativ, wenn man sich die relativ junge Geschichte von Wortlängenforschungen vor Augen hält.

Das Wort ist, ebenso wie der Satz, zentraler Bestandteil eines jeden Prozesses der Textkonstruktion. Ungeachtet dieser zentralen Rolle ist die Wortlänge als eigenständige theoretische Kategorie von der Linguistik und Textwissenschaft lange Zeit vernachlässigt worden. Erst in jüngster Zeit, insbesondere im Zuge des Aufkommens der synergetischen Linguistik, ist die Frage des Häufigkeitsvorkommens von Wörtern einer bestimmten Länge ("Wortlängenhäufigkeiten") in Texten (einer gegebenen Sprache, eines individuellen Autors, eines be-

¹ Address correspondence to: Peter Grzybek, Universität Graz, Institut für Slawistik, Merangasse 70, A-8010 Graz. E-mail: grzybek@uni-graz.at

stimmten Genres, usw.) theoretisch in systematische Kontexte integriert worden, und erst kürzlich ist eine spezifische Theorie der Häufigkeitsverteilung(en) von Wortlängen erarbeitet worden.

Sicherlich hat es, insbesondere in den 50er und 60er Jahren, eine ganze Reihe von Ansätzen gegeben, vor allem im Kontext von Strukturalismus und Informationstheorie, bei denen quantitative Aspekte des Wortes in erster Linie als relevant für stilistische Studien autoren- oder diskursspezifischer Charakteristika angesehen wurden. Diese Studien konzentrierten sich in erster Linie auf die mittlere Wortlänge – ein Vorschlag, der bekannterweise bereits 1851 von dem englischen Mathematiker und Logiker Augustus de Morgan (1806-1871) unterbreitet wurde. Natürlich wurde auch schon in diesen früheren Untersuchungen dem Umstand Rechnung getragen, dass Mittelwerte aufgrund von sehr unterschiedlichen Voraussetzungen zustande kommen, d.h. auf sehr unterschiedlichen Häufigkeitsverteilungen basieren können. Folglich konzentrierte sich auch die sogenannte quantitative Stilistik nicht nur auf Mittelwerte, sondern auch auf Varianzen als spezifische Textcharakteristika. Aus heutiger Sicht sind diesem Vorgehen aber zumindest zwei wesentliche Einwände entgegenzusetzen:

1. Mittelwert und Varianz sind nur zwei spezifische Kenngrößen der Verteilung; ohne Zweifel erlauben diese Maße einen korrekten, dennoch aber nur eingeschränkten Blick auf das gesamte Datenmaterial. Um zu aussagekräftigeren Ergebnissen zu gelangen, ist es deshalb notwendig, zusätzliche charakteristische Kenngrößen (wie z.B. Standardfehler des Mittelwerts, der Standardabweichung, des Medians, ebenso weitere Variations- und Dispersionsmaße wie Variationskoeffizient und dessen Standardfehler, Dispersionsindex und dessen Streuung, Schiefe, Kurtosis und deren Standardfehler, und viele andere mehr) zu verarbeiten. Auch verschiedene auf der Entropie und ihrer Varianz basierende Maße, wie z.B. (relative) Redundanz und Wiederholungsrate, usw. müssen im Detail analysiert werden. Antić/Djuzelic/Grzybek/Stadlober (2003) haben eine Liste entsprechender Kenngrößen erarbeitet, deren Relevanz und Zusammenhänge sowohl theoretisch als auch empirisch auszuloten sein werden.
2. In der Regel hat sich die Analyse von Wortlängen auf die mittlere Wortlänge und/oder Varianz nicht individueller Texte, sondern (mehr oder weniger klar definierter) Textkorpora, sei es eines Autors, einer bestimmten literarischen Periode, einer Gattung, eines spezifischen Funktionalstils o.ä. konzentriert. Ausgehend von der (falschen) Annahme, dass durch die Akkumulation von möglichst vielen Texten (einer Sprache, eines Genres, eines Autors, usw.) so etwas wie eine "Norm" etabliert werden könnte, hat man die Tatsache verdrängt, dass jeder Text das spezifische (individuelle) Ergebnis eines Prozesses der Textgenerierung ist, der durch bestimmte linguistische und/oder psycholinguistische Regularitäten gesteuert wird. Deshalb geht man in der gegenwärtigen Quantitativen Linguistik davon aus, dass ein Textkorpus nichts anderes als eine heterogene Textmischung (ein "Quasi-Text") ist. Daher kann es keine Textakkumulation geben, die so homogen ist, als dass sie durch ein einheitliches Verteilungsmodell beschrieben werden könnte. Für die konkrete Wortlängenforschung bedeutet die Existenz solcher "lokaler" Einflussfaktoren (wie Autorschaft, Gattung, Funktionalstil, usw.) die getrennte Analyse vollständiger Texte, und weder die Analyse von Textkorpora noch von Textauszügen.
3. Keine spezifische Kenngröße (oder eine Kombination von Kenngrößen) erlaubt eine Antwort auf die Frage, wie diese Kenngrößen durch die Häufigkeitsverteilung selbst motiviert sind, d.h. auf die Frage, wie sich die jeweiligen Häufigkeiten der *i*-silbigen Wörter innerhalb einer Verteilung ausnehmen. Bisher verfügbare empirische Ergebnisse zeigen in der Tat, dass die Häufigkeiten, mit der ein-, zwei-, drei- usw. mehrsilbige Wörter in Texten vorkommen, nicht chaotisch, sondern nach bestimmten Regularitäten organisiert sind; die Kenntnis dieser Gesetzmäßigkeiten erlaubt tiefe Einsicht in die Textstruktur und in die Textverarbeitung. Im Gegensatz zu früheren Annahmen, denen zufolge ein einziges, einheitliches Gesetz für die Wortlängenhäufigkeiten verantwortlich sein könnte (Čebanov, Fucks, u.a.), geht man heutzutage von einem flexiblen System eines übergeordneten Basismodells mit verschiedenen Modifikationen aus, die mit spezifischen (sprach-, autoren-, gattungs-, usw. bedingten) Faktoren in Zusammenhang stehen.

Die Frage jedoch, welche Faktoren auf die Wortlänge und deren Häufigkeit in Texten

Einfluss haben, und wie diese Faktoren möglicherweise interagieren, ist bislang noch nie systematisch untersucht worden. Theoretisch bestehen zwei Möglichkeiten, wie solche Einflussfaktoren ins Spiel kommen können:

- a. Gemäß der ersten Annahme führt der Einfluss solcher Faktoren wie Autorschaft, Gattung, Diachronie usw. zu verschiedenen Typen von Häufigkeitsmodellen. Falls diese durchaus plausible Hypothese sich bestätigen sollte, dann wäre es als nächstes wichtig zu wissen, inwiefern sich die für die Texte einer gegebenen Sprache relevanten Modelle gegebenenfalls auf ein gemeinsames, übergeordnetes Modell zurückführen und als dessen Modifikationen interpretieren lassen.
- b. Gemäß der zweiten Annahme führen die genannten Faktoren zu unterschiedlichen Modellen; in diesem Fall würden sich Einflussfaktoren wie die genannten nicht auf das spezifische Häufigkeitsmodell, wohl aber auf die spezifischen Parameter eines gegebenen Modells auswirken.

Beide Varianten lassen sich in der konkreten Textrealität beobachten; dennoch sind noch keine systematischen Untersuchungen durchgeführt worden, die dieser Frage konkret nachgehen. Deshalb konzentriert sich das Projekt in erster Linie auf die folgenden drei Fragebereiche:

- I. Die systematische Untersuchung von Wortlänge und Wortlängenhäufigkeiten in Texten aus drei verschiedenen slawischen Sprachen (Kroatisch, Russisch, Slowenisch) zielt auf die Unterscheidung von sprach(en)-spezifischen und sprach(en)-übergreifenden Faktoren;
- II. Die systematische Untersuchung bestimmter Autorenstile, texttypologischer Besonderheiten usw. zielt auf die Isolierung von Faktoren, die möglicherweise die Wortlänge und deren Häufigkeitsverteilung in Texten beeinflussen; die Relevanz dieser Faktoren und mögliche Zusammenhänge zwischen ihnen wird in einem weiteren Schritt zu untersuchen sein.
- III. Vorausgesetzt, es lassen sich auf die beiden zuletzt genannten Fragen Antworten finden, lässt sich in einem nächsten Schritt die Richtung der Fragestellung umdrehen und danach fragen, ob sich bestimmte individuelle Texte einem bestimmten Autor, einem bestimmten Genre etc. mit einer bestimmten Wahrscheinlichkeit zuordnen lassen. Bescheidener formuliert lautet die Frage: Welchen Beitrag können Wortlängenerforschungen zur Beantwortung dieser Frage(n) beitragen?

Das insgesamt auf drei Jahre ausgerichtete Programm lässt sich in drei aufeinander folgende Phasen untergliedern:

1. Als erstes gilt es, eine umfangreiche Textdatenbank mit jeweils ca. 1000 Texten in jeder der drei zur Diskussion stehenden slawischen Sprachen aufzubauen, einhergehend mit entsprechenden Meta-Daten. Bei der Zusammenstellung dieses Korpus werden insbesondere texttypologische Faktoren zu berücksichtigen sein, damit eine geeignete Basis für die anschließend durchzuführenden statistischen Analysen geschaffen ist. Es wird zu überlegen sein, inwiefern diese Datenbank auch für den externen Gebrauch zugänglich gemacht werden kann, wofür freilich eine geeignete Korpuschnittstelle zu erstellen ist. Weiters ist in dieser Phase geeignete Textanalyse-Software zu erstellen (die nach Möglichkeit zukünftige, auch über slawische Sprachen hinausgehende Möglichkeiten der Weiterentwicklung vorsieht).
2. Im nächsten Schritt kommt es darauf an, die Texte für die anstehenden Analysen aufzubereiten; diese ausschließlich textbezogene Präparation beinhaltet neben einer einheitlichen Behandlung von Abkürzungen, Überschriften, Zahlen, Fremdwörtern, u.a. auch Fragen der Satzdefinition, der Unterscheidbarkeit von narrativen, deskriptiven, dialogischen und anderen Sequenzen. Im Anschluss an diese Textaufbereitungen lassen sich mit den in der ersten Phase erstellten Analyseprogrammen erste statistische

Auswertungen durchführen, die für jeden Text die entsprechenden Rohdaten liefern, die in für die weiterführenden statistischen Analysen geeignete Datenfiles abzuspeichern sind.

- Die letzte Phase beinhaltet die quantitativen und qualitativen Auswertungen. In diesem Abschnitt wird es darauf ankommen, ein passendes Verteilungsmodell (oder, in Abhängigkeit von den Ergebnissen, mehrere geeignete Modelle) zu finden. Weitere Analysen werden sich dann auf die Faktoren richten, die entweder die konkreten Parameter des Verteilungsmodells oder aber den Verteilungstyp insgesamt beeinflussen. Dieser Arbeitsschritt wird zur Anwendung von Diskriminanz- und Clusteranalysen führen, was tiefe Einsicht in Fragen der Texttypologie, Textklassifikation und Textdiskrimination verschaffen sollte.

Wenn sich in der Tat relevante Regularitäten beobachten lassen sollten, so werden diese auch allgemein für Prozesse der Informationsverarbeitung als relevant anzusehen sein. In diesem Falle stellen die angewendeten statistischen Verfahren und Methoden eine optimale Basis für weiterführende interdisziplinäre Studien dar, denen es darauf ankommt, die vermeintliche Kluft zwischen den "zwei Kulturen" von Natur- und Kulturwissenschaften zu überbrücken.

2. Das Grazer Projekt: Phase I (2002-2003)

Die folgende Darstellung hat zum Ziel, das im ersten Projektjahr (1.4.2002 – 31.3.2003) Erreichte vor dem Hintergrund der oben vorgestellten Projektziele zu referieren.

2.1. Erstellung der Datenbank und Entwicklung von Analyseprogramme

Wie oben dargestellt, ist es in der ersten Projektphase ein zentrales Ziel gewesen, die *systematische* und *automatisierte* Untersuchung der Wortlängen(häufigkeiten) in den erwähnten slawischen Sprachen vorzubereiten.

In diesem Zeitraum wurde neben der Recherche nach elektronisch verfügbaren kroatischen, slowenischen und russischen Texten (bzw. dem zusätzlichen Einscannen weiterer Texte) die einheitliche unicode-fähige Codierung der Texte (CP 1250 bzw. CP 1251) bewältigt. Nach derzeitigem Stand wurde die folgende Anzahl von Texten (Analyseeinheit: "Inhaltlich abgeschlossene Einheit") in einer zentralen Datenbank gespeichert und zur weiteren Bearbeitung vorbereitet:

Sprache	Anzahl von Texten
Kroatisch	1050
Russisch	1303
Slowenisch	1290
Gesamt	<u>3643</u>

Für die systematische Verwaltung der einzelnen Texte wurde ein Linux-Server eingerichtet, wobei darauf geachtet wurde, strukturell zwischen Daten (einzelne Texte, abgespeichert als Files) und Metadaten (Informationen über die Autoren, Titel, Textquelle, Texttyp, abgelegt in einer POSTGRESQL-Datenbank) zu trennen. Das verwendete Metadaten-Schema wurde unter anderem daraufhin optimiert, Textgruppen nach bestimmten Kriterien aus dem Korpus zu

extrahieren. Um die Administration der Textdatenbank (Speicherung der einzelnen Texte, Eingabe und Verwaltung der Metadaten) zu erleichtern, wurde ein projektintern zugänglicher Web-Zugang implementiert, der den Zugriff auf die Daten von beliebigen Orten aus garantiert. Dieses Verfahren sieht bereits die Möglichkeit späterer externer Zugänge vor.

Da keine Analyse-Software zur Bestimmung der Wortlänge in Texten und die daraus notwendigerweise abzuleitenden statistischen Kenngrößen verfügbar ist, wurde innerhalb des Projektes eine eigene Software auf der Basis der Programmiersprache PERL entwickelt. Geplant ist, dass bei endgültiger Fertigstellung folgende Analyseschritte automatisiert durchgeführt werden können:

- a.) Statistische Auswertung von Graphemsystemen (Vorkommenshäufigkeit);
- b.) Auswertungen der Silbenstruktur;
- c.) Automatische Bestimmung der Wortlänge aufgrund unterschiedlicher Wortdefinitionen und Maßeinheiten (Graphem, Silbe).

In einem weiteren Schritt werden diese – auf kroatische, russische und slowenische Texte ausgerichteten – Analyseprogramme direkt auf dem Server ausgeführt und garantieren somit eine bedienerfreundliche und effiziente statistische Textanalyse. Vorbereitet ist sowohl die Möglichkeit der lokalen (externen) Applikation als auch der Erweiterung auf andere Sprachen.

2.2. Theoretische Grundlagenarbeiten

2.2.1. Wortlängenforschungen im Kontext der Geschichte der Quantitativen Linguistik

Mit dem Ziel, die spezifischen Untersuchungen zur Wortlänge bzw. zu Wortlängenhäufigkeiten in den allgemeinen Kontext der Quantitativen Linguistik und Quantitativen Textanalyse einzuordnen, wurden eine Reihe synoptischer Studien zu diesem Bereich durchgeführt. So geht Grzybek (2003c) detailliert auf die Frage der Geschichte der Wortlängenforschungen ein: Neben der chronologischen Aufarbeitung der einzelnen Analysen seit Ende des 19. Jahrhunderts finden sich Darstellungen bis hin zu den neuesten Ansätzen im Lichte einer synergetischen Linguistik, in der die Evolution der in Betracht zu ziehenden statistischen Modelle für Wortlängen näher untersucht werden.

Ausgehend von dieser konkreten Fragestellung wurden zwei weitere Überblicksartikel zum Status und zur Entwicklung der quantitativen Linguistik in Russland (bzw. der Sowjetunion) verfasst, mit denen eine Einbettung in einen breiteren Forschungskontext vorgenommen werden kann. So beschäftigen sich Grzybek/Kelih (2003a) mit den Anfängen quantitativer Sprachanalysen in Russland seit Mitte des 19. Jahrhunderts bis hin zu den quantitativ orientierten Studien im Umfeld des russischen Formalismus. Daran anknüpfend beschäftigen sich Grzybek/Kelih (2003b) mit dem neuerlichen Start der quantitativen Linguistik in Russland nach 1956 im Kontext von strukturalistischen, kybernetischen, und semiotischen Untersuchungen bzw. im Lichte der maschinellen Übersetzung, um sodann den Bogen zu den aktuellen Forschungen der russischen quantitativen Linguistik zu spannen.

2.2.2. Tagung zur Quantitativen Textanalyse (Graz, 21.-23. Juni 2002)

Unmittelbar nach Projektbeginn wurde vom 21. bis 23. Juni 2002 eine international besetzte Konferenz unter dem Titel "Wortlängen in Texten. Internationales Symposium zur quanti-

tativen Textanalyse" veranstaltet. Ziel dieser interdisziplinär ausgerichteten Veranstaltung war es, ein Forum für die Präsentation von aktuellen Forschungsergebnissen auf dem Gebiet der quantitativen Linguistik zu schaffen. Dieses sollte die Möglichkeit bieten, mit führenden VertreterInnen der quantitativen Linguistik bereits zu Projektanfang die Untersuchungen zur Wortlängen(häufigkeiten) in slawischen Sprachen auf breiter Ebene zu diskutieren und die Perspektiven bzw. die inhaltliche Richtung einschlägiger Forschungen zu koordinieren.

Da Verlauf und Ergebnisse dieser Veranstaltung andernorts im Rahmen zweier Konferenzberichte vorgestellt wurden (Kelih/Grzybek 2002, Antić/Kelih/Grzybek 2002), kann hier auf eine eingehende Darstellung verzichtet werden. Anzumerken bleibt, dass ein Großteil der bei diesem Symposium gehaltenen Beiträge unter dem Titel "Word Length Studies and Related Topics" erscheint (Grzybek ed. 2003).

2.2.3. Wortlängenforschungen – Fragen der Wort-Definitionen

In unmittelbarer Relevanz für die Wortlängenforschung wurden im ersten Projektjahr vor allem Fragen der Definition und Quantifizierung der Einheit ‚Wort‘ in einer Reihe von Arbeiten reflektiert und diskutiert. Bekanntlich ist für quantitative Untersuchungen in jedem Fall eine exakt nachvollziehbare Definition der zu untersuchenden Einheiten notwendig, wobei auf unterschiedliche theoretische Konzeptionen der Linguistik zurückgegriffen werden kann.

Ausgehend von einer *graphematischen* Konzeption des Wortes, derzufolge ein Wort eine in (schriftlichen) Texten durch Leerstellen abgegrenzte Einheit darstellt, wurde diese Konzeption durch eine *graphematisch-phonologische* Definition erweitert: Wird die Wortlänge in slawischen Sprachen in der Anzahl der Silben pro Wort bestimmt, ergibt sich bei der Bestimmung der Wortlänge aufgrund des graphematischen Kriterium eine als problematisch anzusehende Gruppe von "0-silbigen" Wörtern (Gruppe von Präpositionen, die aus phonologischer Sicht als Enklitika behandelt werden können). Diskutiert wurde diese Problematik im Rahmen einer empirischen Studie unter dem Titel: "On the question of so-called 0-syllable words in determining word length" (vgl. Antić/Kelih 2003). Hierbei ging es vor allem um die Auswirkungen dieser beiden unterschiedlichen Wortdefinitionen auf übliche statistische Kenngrößen.

Auf der Basis dieser Ergebnisse wurde in einer weiteren Studie auf ähnliche Art und Weise eine andere Konzeption des Wortes diskutiert, und zwar im Hinblick auf die Implikationen einer *phonologischen* Wortdefinition (Taktgruppen) in Texten. In einer entsprechend angelegten vergleichend-statistischen Analyse der Wortlänge aufgrund von drei unterschiedlichen Konzeptionen (graphematisch, graphematisch-phonologisch, phonologisch) konnte anhand der Wortlängen in russischen Texten die systematische Verschiebung von statistischen Kenngrößen (s. Kelih/Grzybek 2003a) gezeigt werden.

Zusammengefasst zeigt es sich, dass die Bestimmung der Wortlänge unter Berücksichtigung von unterschiedlichen linguistischen Konzeptionen stringent vollzogen werden kann. Insgesamt stellt es sich heraus, dass die Wahl von unterschiedlichen Wortdefinitionen zwar – wie nicht anders zu erwarten – zu unterschiedlichen quantitativen Größen führt, dass diese jedoch eine *systematische Verschiebung* bewirken (was in letzter Konsequenz eine wechselseitige Transformation der Ergebnisse unabhängig von der jeweils gewählten Basiseinheit erlaubt).

2.2.4. Frage der konstituierenden Elemente des Wortes

Im Zusammenhang mit der Definition und Bestimmung der Wortlänge ist die Frage der konstituierenden Elemente des Wortes von unmittelbarer Relevanz. Denn bei der Berechnung der Wortlänge ist prinzipiell die Wahl unterschiedlicher konstituierender Einheiten möglich, die jeweils – hierarchisch gesehen – auf verschiedenen Ebenen anzusetzen sind: so ist ein Wort z.B. messbar in Graphemen, Phonemen, in Silben oder in Morphemen.

In diesem Zusammenhang wurde die grundsätzliche Frage der statistischen Modellierung von Graphemsystemen untersucht (vgl. Grzybek/Kelih 2003a,b); als ein wesentliches Ergebnis konnte festgestellt werden, dass es sich – zumindest bei den im Projekt zu berücksichtigenden Sprachen – beim graphematischen System um ein äußerst stabiles System handelt, und dass Einflussfaktoren (wie Textlänge, Textgattung, Stichprobengröße usw.) offensichtlich keine wesentliche modifizierende Rolle spielen.

Neben der prinzipiellen Frage der Wortdefinition (vgl. 2) wird weiteres zu überprüfen sein – wie eingangs erwähnt – welche Auswirkungen unterschiedliche Maßeinheiten (Graphem, Silbe, Morphem) auf die Bestimmung der Wortlänge zeigen. Zu überprüfen wird sein, inwiefern – ebenso wie für unterschiedliche Wortdefinitionen (s.o.) – *systematische Verschiebungen* nachgewiesen werden können (vgl. Kelih 2003b, Kelih/Grzybek 2003b).

2.2.5. Fragen von Einflussfaktoren

Bereits im Vorfeld systematischer Analysen der Wortlängen(häufigkeiten) in Texten slawischer Sprachen ist es von Bedeutung, unterschiedlichste Einflussfaktoren auf die Wortlänge in Erwägung zu ziehen.

Nicht nur im Hinblick auf die beabsichtigten Datenanalysen, sondern vor allem auch in Anbetracht der Datenbankstruktur (werden ganze Romane oder einzelne Romankapitel als ‚Text‘ definiert?) ist die Frage der Datenhomogenität von zentraler Bedeutung. Strauss/Grzybek/Altmann (2003) haben bei der Untersuchung des Zusammenhangs von Wortfrequenz und Wortlänge Argumente dafür bereitgestellt, Texte ausschließlich auf der Ebene homogener Texteinheiten zu berücksichtigen; in einer sich daraus ergebenden Folgestudie haben Grzybek/Altmann (2003) darüber hinaus in einer methodologisch ausgerichteten Untersuchung auf die Frage der Datenaufbereitung für quantitative Analysen aufmerksam gemacht.

Neben Faktoren wie Autorschaft, Zeitraum der Entstehung der jeweiligen Texte, u.ä. wird besondere Rücksicht auf die Frage des Texttyps zu legen sein. In ausführlicher Weise wird die Frage einer quantitativen Texttypologie – auch in Hinblick auf die strukturelle Gliederung der zu analysierenden Texte in der dazu angelegten projektinternen Datenbank – aus theoretischer Sicht in Grzybek (2003b) diskutiert. Entsprechende empirische Untersuchungen zu diesem Fragenkomplex – auf der Basis von slowenischen Texten – werden in Kelih (2003a) durchgeführt.

Abgesehen von den theoretischen Begründungen wurden eine Reihe vorbereitender Untersuchungen im Hinblick auf die Frage der Texttypologie als einer möglichen Einflussvariable durchgeführt.

Wie in einer Untersuchung von Grzybek/Stadlober (2003) zur Frage der Wortlänge in tschechischen Texten von Karel Čapek gezeigt werden konnte, sprechen die erhaltenen Ergebnisse dafür, dass offensichtlich weniger die Autorschaft oder der Zeitpunkt der Entstehung der analysierten literarischen Werke eine Rolle spielt, als vielmehr der – ebenfalls notwendigerweise einer Definition unterliegende – Texttyp.

In diesem Zusammenhang war es unter anderem auch von Bedeutung, die Anwendbarkeit der eingeschlagenen Untersuchungsdesigns auch in spezifischen Textsorten zu prüfen. In Be-

zug auf diese Frage liegen zwei Detailstudien von Grzybek (2002, 2003a) zu poetischen (metrisch gebundenen) Verstexten von A.S. Puškin sowie eine weitere Untersuchung an Sprichwort-Material (Grzybek 2003d) vor, die als paradigmatische Basis für weitere Studien dienen können.

2.2.6. Vorbereitung der Analyse statistischer Textgrößen

Abgesehen von den im ersten Projektjahr gelösten textwissenschaftlichen und programmier-technischen Fragestellungen wurde eine Reihe von statistisch-theoretischen Arbeiten durchgeführt.

So konnten Stadlober/Djuzelic (2003) im Rahmen der oben angeführten Konferenz der Frage nachgehen, inwiefern – unter Einsatz von multivariaten Verfahren der Statistik – bestimmte aus der Wortlänge ableitbare statistische Kenngrößen als potentielle Variablen für eine quantitative Texttypologie eingesetzt werden können. Ausgehend von diesen methodologischen Überlegungen wird eine Ausarbeitung von relevanten statistischen Kenngrößen zur Beschreibung von Häufigkeitsverteilungen, die unmittelbar aus der Wortlänge her- und ableitbar sind (vgl. Antić/Djuzelic/Grzybek/Stadlober 2003). Selektiv aufgearbeitet sind einige dieser Kenngrößen in Djuzelic (2002), wo unter anderem statistische Kenngrößen der Wortlänge als Diskriminationsfaktoren für automatische Textklassifizierungen verwendet werden.

Abschließend sei angemerkt, dass die erwähnten statistischen Kenngrößen in die oben angeführte Analyse-Software eingebaut werden sollen, um so eine effiziente und systematische Analyse der Wortlänge und den damit in Zusammenhang stehenden Einflussfaktoren in den nächsten zwei Projektjahren ermöglichen werden.

2.2.7. Internationale Kooperationen

Im Rahmen der bisherigen Projektarbeit verstärkt sich derzeit die Zusammenarbeit auf internationaler Ebene. Abgesehen von einzelnen Kontakten in verschiedene europäische Länder kristallisiert sich insbesondere eine verstärkte Zusammenarbeit mit einer Reihe von slowakischen Institutionen heraus. Konkret handelt es sich um das Institut für Mathematik der Slowakischen Akademie der Wissenschaften Bratislava (Gejza Wimmer), das Institut für Mathematik der Matej Bela Universität Banská Bystrica (Jana Kusendová), sowie das Institut für Slowakische Sprache der Universität Trnava (Emília Nemcová). Um die sich abzeichnenden Kooperationen auf institutionalisierter Ebene realisieren zu können, wird derzeit ein Austausch über verschiedene europäische Förderprogramme eingeleitet.

Literatur

(Die folgenden Angaben beziehen sich ausschließlich auf die im Grazer Projekt durchgeführten Arbeiten und sollen keinen Literaturbericht zum Thema darstellen)

- Antić, G.; Djuzelic, M.; Grzybek, P.; Stadlober, E.** (2003). *Statistische Kenngrößen zur Beschreibung von Wortlängenhäufigkeitsverteilungen*. [Ms.]
- Antić, G.; Kelih, E.; Grzybek, P.** (2002). Word Length in Texts. An International Symposium on Quantitative Text Analysis. In: *Journal of Quantitative Linguistics* [Im Druck].
- Antić, G.; Kelih, E.** (2003). On so-called 0-syllable words in determining word length. In: Grzybek, P. (ed.), *Word Length Studies and Related Topics*. [In print]

- Djuzelic, M.** (2002): Einflussfaktoren auf die Wortlänge und ihre Häufigkeitsverteilung am Beispiel von Texten slowenischer Sprache. Technische Universität Graz, Diplomarbeit. [<http://www.cis.tugraz.at/stat/dthesis/djuz02.zip>]
- Grzybek, P.** (2002). Versuchen wir einmal, die Kräfte aus dem Gleichgewicht zu bringen... Quantitative Aspekte von Puškins *Evgenij Onegin* und *Domik v Kolomne*. In: J. Bernard; P. Grzybek; A. Pokrivčák; G. Withalm (eds.), *Form – Struktur – Komposition. Pragmatik und Rezeption*. Wien, 305-335. [= Special Issue of: *Semiotische Berichte*, 26,1-4.]
- Grzybek, P.** (2003a). Quantitative Aspekte slawischer Texte (am Beispiel von Puškins *Evgenij Onegin*. (Beitrag zum XIII. Internationalen Slawistenkongress, Ljubljana 2003.) In: *Wiener Slawistisches Jahrbuch*, 49. [Im Druck]
- Grzybek, P.** (2003b). Empirische Texttypologie in quantitativer Sicht. [In Vorb.]
- Grzybek, P.** (2003c): History and Status of Word Length Frequency Studies. In: Grzybek, P. (ed.) (2003). *Word Length Studies and Related Topics*. [In print]
- Grzybek, P.** (2003d): Zur Wortlänge und ihrer Häufigkeitsverteilung in Sprichwörtern (Am Beispiel slowenischer Sprichwörter, mit einer Re-Analyse estnischer Sprichwörter). In: Palm-Meister, Christine (Hg.), *Europhras 2000*. Tübingen. [Im Druck]
- Grzybek, P.; Altmann, G.** (2003): Oscillation in the frequency-length relationship. In: *Glottometrics* 5, 97-107.
- Grzybek, P.** (ed.) (2003). *Word Length Studies and Related Topics. Proceedings of the Graz Conference on Quantitative Text Analysis, June 21-23, 2002*. [In print]
- Grzybek, P.; Kelih** (2003a): Zu den Anfängen der quantitativen Linguistik in Russland. *International Handbook of Quantitative Linguistics*. [In print]
- Grzybek, P.; Kelih, E.** (2003b). Quantitative Linguistik in Russland seit 1956. *International Handbook of Quantitative Linguistics*. [In print]
- Grzybek, P.; Kelih, E.** (2003a). Häufigkeit russischer Grapheme. Teil I: Zur Geschichte der Untersuchung russischer Graphemhäufigkeiten. In: Deutschmann, P.; Höller, H. (eds.), *Datenverarbeitung in Sprach-, Literatur- und Kulturwissenschaft*. Graz. [Im Druck].
- Grzybek, P., Kelih, E.** (2003b). Häufigkeit russischer Grapheme. Teil II: Modelle von Häufigkeitsverteilungen. In: Deutschmann, P.; Höller, H. (eds.), *Datenverarbeitung in Sprach-, Literatur- und Kulturwissenschaft*. Graz. [Im Druck].
- Grzybek, P., Stadlober, E.** (2002): The Graz Project on Word Length (Frequencies). In: *Journal of Quantitative Linguistics*, 9; 187-192.
- Grzybek, P.; Stadlober, E.** (2003). Zur Prosa Karel Čapeks – Einige quantitative Bemerkungen. In: Kempgen, S. (ed.), *Sprach- und Textstudien*. [Im Druck]
- Kelih, E.** (2003a): Empirische Texttypologie aus quantitativer Sicht. [In Vorb.]
- Kelih, E.** (2003b): Wortlänge in Silben und Morphemen (am Beispiel des Russischen). [In Arbeit]
- Kelih E.; Grzybek, P.** (2002). Wortlängen in Texten. Internationales Symposium zur quantitativen Textanalyse. In: *etc. Empirische Text- und Kulturforschung / Empirical Text and Culture Research*, 2; 89-91.
- Kelih, E.; Grzybek, P.** (2003a): Wortdefinitionen und Wortlängenforschung. [Im Druck]
- Kelih, E.; Grzybek, P.** (2003b): Wortlänge in Silben und Graphemen (am Beispiel des Russischen). [In Arbeit]
- Stadlober, E., Djuzelic, M.** (2003): Multivariate Statistical Methods of Quantitative Text Analysis. In: Grzybek, P. (ed.) (2003). *Word Length Studies and Related Topics*. [In print]
- Strauss, U.; Grzybek, P.; Altmann, G.** (2003): The more the better? Word Length and Word Frequency. [Submitted]

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitschrift für die quantitative Erforschung von Sprache und Text

Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden

Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals

Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors

Glottometrics can be downloaded from the **Internet**, obtained on **CD-ROM** (in PDF) or in form of **printed copies**

Herausgeber – Editors

G. Altmann	02351973070-0001@t-online.de
K.-H. Best	kbest@gwdg.de
P. Grzybek	grzybek@uni-graz.at
L. Hřebíček	hrebicek@orient.cas.cz
R. Köhler	koehler@uni-trier.de
V. Kromer	applied@newmail.ru
O. Rottmann	otto.rottmann@t-online.de
A. Schulz	reuter.schulz@t-online.de
G. Wimmer	wimmer@mat.savba.sk
A. Ziegler	arneziegler@compuserve.de

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM's or printed copies to

RAM-Verlag RAM-Verlag@t-online.de

Herunterladen / Downloading: <http://www.ram-verlag.de>

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Glottometrics. –6 (2003) –. – Lüdenscheid: RAM-Verl., 2003

Erscheint unregelmäßig. – Auch im Internet als elektronische Ressource unter der Adresse <http://www.ram-verlag.de> verfügbar.-

Bibliographische Deskription nach 6 (2003)

ISSN 1617-8351

Contents

Hřebíček, L. Some aspects of the power law	1-8
Best, K.-H. Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes	9-34
Wilson, A. Word-length distribution in modern Welsh prose texts	35-39
Dshurjuk, T.V., Levickij, V.W. Satztypen und Satzlängen im Funktional- und Autorenstil	40-51
Rottmann, O. Word length in the Baltic languages – are they of the same type as the word lengths in the Slavic languages?	52-60
Strauss, U., Altmann, G. Age and polysemy of words	61-64
Wheeler, E.S. Multidimensional scaling to visualize text separation	65-69
Jüngling, R., Altmann, G. Python for linguistics?	70-82
Popescu, Ioan-Iovitz On a Zipf's Law extension to impact factors	83-93
Project report Kelih, E., Grzybek, P., Stadlober, E. Das Grazer Projekt zu Wortlängen(häufigkeiten)	94-102
History of Quantitative Linguistics I. V.J. Bunjakovskij (by P. Grzybek) II. B. Trnka – The first bibliography (by L. Uhliřová)	103-106
Books received	107