Anzeiger für Slavische Philologie
(XXXIX) 2011, 39–74

Der Satz und seine Beziehungen

I: Satzlänge und Wortlänge im Russischen (Am Beispiel von L.N. Tolstojs «Анна Каренина»)

Peter Grzybek (Graz)

1. Satzlänge

Untersuchungen des Satzes und seiner Länge sind in den Bereichen der Sprach-, Literatur- und Textwissenschaften wiederholt mit den unterschiedlichsten Fragestellungen verbunden und im Rahmen eines breiten Forschungsspektrums verankert worden. Angefangen von der Satzlänge als einem autor- oder stilspezifischen Charakteristikum bis hin zu Fragen der Textverständlichkeit wurde die Satzlänge von Texten immer wieder als ein wesentliches Kriterium der Textgestaltung angesehen.

Auch wenn die Geschichte der Satzlängenforschung mit den Arbeiten von Sherman (1888) auf eine mehr als 100-jährige Geschichte zurückblickt, kann und soll es an dieser Stelle nicht um eine ausführliche Darstellung der entsprechenden Forschungsgeschichte gehen. Bestenfalls lässt sich hier sozusagen „en passant“ auf die bekannte und für den russischen Bereich nach wie vor als paradigmatisch angesehene Serie von Arbeiten von Leskiss (1962, 1963, 1964) verweisen, in denen vor allem die Frage der Häufigkeitsverteilung von Satzlängen – also die Frage: wie oft kommen Sätze einer gegebenen Länge im jeweils untersuchten Material vor – im Vordergrund stand. Nachdem Forscher wie Yule (1939) oder Williams (1939) bereits Ende der 30er Jahre auf die von ihnen vermutete autorenspezifische Dimension der Satzlänge hingewiesen hatten, untersuchte Leskiss in den genannten Arbeiten an für damalige Zeiten enorm umfangreichem Material – immerhin handelte es sich um ca. 1,4 Millionen Wörter und fast 100.000 Sätze – darüber hinausgehend in textsorten- und funktionalstil-differenzierten Studien weitere innersprachliche Faktoren, die er als mit der Satzlänge in Beziehung stehend herausarbeitete. Eine Fortführung dieser Forschungsrichtung findet sich in jüngerer Zeit vor allem in Arbeiten wie denen von Roukk (2001, 2005) und vor allem Kelih (2002) sowie Kelih et al. (2006).

Auch wenn die zuletzt genannten Arbeiten sich bereits im Kontext der Quantitativen Linguistik verorten, die Frage der Satzlängenhäufigkeitsverteil-

lung auf einer theoretischen Modellierungsebene behandeln und auch textsortenspezifische Unterschiede im Rahmen einer bzw. mit dem Ziel einer quantitativen Typologie untersuchen, konzentrieren sie sich in ihrem Fokus nach wie vor auf die Häufigkeitsverteilung und damit gewissermaßen auf die isolierte Ebene des Satzes und seiner Länge. Zusammenhänge der Satzlänge zu Einheiten und Konstrukten anderer sprachlicher Ebenen innerhalb des jeweils gegebenen Textes bzw. Textmaterials sind im Vergleich dazu im Hinblick auf das Russische nur spärlich untersucht. Dies betrifft ebenso die „vertikalen" Beziehungen der Satzlänge zur Länge von Einheiten und Konstrukten anderer sprachlicher Ebenen – d.h. Beziehungen „nach unten" wie etwa zwischen der Satzlänge und der Länge der sie konstituierenden Teilsätze bzw. Wörter – wie die Beziehungen „nach oben" wie etwa zwischen der Satzlänge und der Länge größerer Texteinheiten wie z.B. der Abschnitts- oder der Kapitellänge.

Systematische Untersuchungen in beide Richtungen sind für das Russische Mangelware. Im Hinblick auf Beziehungen zu „höheren" Ebenen gilt dies freilich nicht nur für das Russische; vielmehr stellen entsprechende Untersuchungen insgesamt in der Forschung ein absolutes Desiderat dar. Versuche in dieser Richtung, die im Vorfeld des Russischen Formalismus bereits in den 10er Jahren des 20. Jhds. postuliert, aber nie (und erst recht nicht systematisch) umgesetzt wurden (vgl. Grzybek 2012a), finden sich etwa bei Hřebíček (1995) oder Neumann (2009), und Grzybek (2012b) stellt erstmals eine systematische Beziehung zwischen Satzlänge und Kapitellänge her.

Im Vergleich dazu sind Beziehungen der Satzlänge zu „niedrigeren" Sprachebenen und deren Einheiten bzw. Konstrukten zumindest im Hinblick auf andere Sprachen hinreichend untersucht worden, und es liegen auch erste Analysen für das Russische vor – vgl. Roukk (2007), Grzybek et al. (2008).

Diese Studien sind in erster Linie vor dem Hintergrund des bekannten Menzerath-Altmann-Gesetzes (Altmann 1980, Altmann/Schwibbe 1989, Cramer 2005) durchgeführt worden, welches im Wesentlichen besagt, dass die ein sprachliches Konstrukt konstituierenden Einheiten umso kürzer sind, je länger das betreffende Konstrukt ist.[1] Die diesem Gesetz zugrunde liegenden theoretischen Annahmen – die spezifische nicht-lineare Relationen zwischen Konstrukt und Konstituenten hypostasieren, auf die unten noch detaillierter

[1] In dieser Form ist das Menzerath-Altmann-Gesetz als *intratextuell* konzipiert zu verstehen, d.h. es ist auf die Verhältnisse innerhalb des jeweils gegebenen sprachlichen Materials (eines Textes, Korpus, usw.) ausgerichtet. Es muss strikt vom Arens-Altmann-Gesetz (Altmann 1983) unterschieden werden, das zwar auf analogen Annahmen basiert, aber *intertextuell* angelegt ist, d.h. auf die Durchschnittsverhältnisse in mehreren Texten (bzw. Korpora) ausgerichtet, deren Werte in einen Mittelwertvektor überführt werden (vgl. Grzybek et al. 2007)

einzugehen sein wird – sind in einer Studie von Roukk (2007) ansatzweise auch auf das Russische übertragen worden. Aufgrund ihrer Analysen, in der die Abhängigkeit der Teilsatzlänge von der Satzlänge untersucht wird, kommt Roukk (2007) jedoch zu der Schlussfolgerung, dass die Ergebnisse ihrer Studie nicht im Einklang mit den Resultaten zu vielen anderen Sprachen zu stehen scheinen, und dass das Menzerath-Altmann-Gesetz nicht für das Russische gültig sei (ebd., 609); auf diese Arbeit und die daraus gezogene Schlussfolgerung wird unten noch ausführlicher einzugehen sein. In Anzweiflung dieser Schlussfolgerung haben Grzybek et al. (2008) erste Untersuchungen zur Abhängigkeit der Wortlänge von der Satzlänge durchgeführt. Dabei hat sich herausgestellt, dass die Verhältnisse in verschiedenen Textsorten vermutlich unterschiedlich sind und einer differenzierten Betrachtung bedürfen, dass jedoch durchaus bekräftigende Evidenz für die Gültigkeit des Menzerath-Altmann-Gesetzes auch im Russischen vorliegt. Allerdings mussten von Grzybek et al. (2008) aufgrund einer Reihe von Beobachtungen bei der Auswertung der Analysen eine Reihe von Einschränkungen in Kauf genommen werden, die im weiteren Verlauf dieser Arbeit noch ausführlicher zur Sprache kommen werden, und die im vorliegenden Text einerseits aufgrund umfangreicheren Analysematerials, andererseits aufgrund von theoretischen Verallgemeinerungen überwunden werden sollen.

Zu diesem Zweck soll im vorliegenden Beitrag am Beispiel des Romans *Анна Каренина* von L.N. Tolstoj die Beziehung der Satzlänge zur Wortlänge detailliert untersucht und in einen entsprechenden theoretischen Zusammenhang gestellt werden, der die isolierte Betrachtung der Satzlänge zugunsten einer systematischen Betrachtungsweise aufgibt, so dass die Satzlänge in ihrer (intratextuellen) dynamischen, synergetischen und selbst-regulativen Relation zur Wortlänge in den Fokus rückt.

Bevor dies im weiteren Verlauf des vorliegenden Texts konkret umgesetzt werden wird, wenn nämlich die funktionalen Zusammenhänge zwischen beiden Größen untersucht und theoretisch modelliert werden (4), scheint es angebracht, zuvor im Anschluss an einleitend zu leistende operationale Definitionen der entsprechenden sprachlichen Einheiten und Konstrukte – also des Satzes (2.1) und des Worts (2.2), – die Häufigkeit von Satz- und Wortlänge in diesen Texten in ihrer „eigenständigen“ Frequenzsystematik nachzuweisen, was es schließlich erlaubt, resümierende Schlussfolgerungen im Hinblick auf den Zusammenhang von Wort- und Satzlänge allgemein und die vermeintliche Sonderstellung des Russischen im Besonderen zu ziehen (5).

2. Operationale Definitionen

In der mit quantifizierenden Methoden arbeitenden Sprach- und Textanalyse beinhaltet einer der ersten Arbeitsschritte obligatorisch eine stringente und nachvollziehbare, auf linguistischer Basis beruhende qualitative Bestimmung der jeweiligen Untersuchungs- und Maßeinheiten. Dies ist im Sinne einer Explikation dessen zu verstehen, *was* im Sinne einer Objektdefinition gemessen wird und *wie* gemessen wird, d.h. in welchen Einheiten die Messung jeweils vorgenommen wird (vgl. dazu die theoretischen Implikationen der Quantifizierung und Formalisierung von Altmann/Lehfeldt 1980). In unserem Fall betrifft dies die sprachlichen Einheiten ‚Satz' (2.1.) und ‚Wort' (2.2). Während es in technischer Hinsicht im Grunde genommen primär um die Frage geht, ob und wie sich aufgrund von objektivierbaren bzw. intersubjektiv nachvollziehbaren Kriterien die (linguistische) Einheit ‚Wort' und ‚Satz' in Texten und Korpora automatisiert bestimmen lässt, leiten sich aus diesem Ausgangsproblem weitere Fragestellungen ab, die mit der Auswirkung der jeweiligen Entscheidung in unmittelbarem Zusammenhang stehen, nämlich z.B.:

a) Sind in Abhängigkeit von der Wahl einer bestimmten Wort- bzw. Satzdefinition im jeweils untersuchten sprachlichen Material statistisch signifikante Unterschiede in der durchschnittlichen Länge dieser sprachlichen Einheiten zu beobachten?
b) Hat die Wahl der Wort- bzw. Satzdefinition einen Einfluss auf postulierte theoretische Häufigkeitsverteilungen und Zusammenhänge?

Es würde zu weit und über den hier gegebenen Rahmen hinaus führen, diesen Fragestellungen innerhalb der hier vorliegenden Studie nachzugehen.[2] Stattdessen soll es um die Erarbeitung linguistisch plausibler und zugleich operationaler Definitionen gehen, die als Grundlage der ins Auge gefassten Untersuchung dienen können.

2.1. Satz: Satzdefinition und automatische Bestimmung des Satzes im Russischen

Die Frage, was genau unter einem ‚Satz' zu verstehen ist, nimmt im Bereich der Syntaxforschung im Allgemeinen und der Satzlängenforschung im Besonderen einen durchaus prominenten Platz ein, wobei die für verschiedene theoretische und praktische Fragestellungen jeweils notwendigen Definitionen von ‚Satz' keineswegs einheitlich sind. So besteht nicht nur innerhalb

[2] Immerhin aber gibt es zu ersten der beiden Fragestellungen eine systematische Studie zum Slowenischen, auf die an dieser Stelle verwiesen werden kann (Kelih/Grzybek 2005).

der quantitativen bzw. statistischen Linguistik relativ große Einigkeit darüber, dass Interpunktionszeichen die Funktion haben, einen schriftlichen Text in Einheiten zu gliedern; von daher können Interpunktionszeichen herangezogen werden, um u.a. auch Satzgrenzen in Texten zu bestimmen. Dies entspricht einer der möglichen Definitionen, die zahlreichen analytischen Arbeiten zugrunde liegt, vertreten etwa von Bünting/Bergenholtz (1995: 27), denen zufolge Sätze in Texten als „Einheiten, die durch Satzzeichen eingegrenzt sind", anzusehen sind. In einer ersten Annäherung handelt es sich hierbei um eine durchaus sinnvolle Operationalisierung der Einheit ‚Satz'. Ungeachtet allfälliger Fragen der Interpunktion in Abhängigkeit von spezifischen Editionsproblemen[3] ist im hier gegebenen Zusammenhang die Frage nach dem Inventar von Interpunktionszeichen, welche genau als Markierung eines expliziten Satzendes herangezogen werden, weitaus wichtiger.

Dabei liegt es auf der Hand, dass eine etwaige Satzdefinition nicht zwangsläufig allgemein sprachübergreifend gültig sein muss. Es können jedoch auch innerhalb einer Sprache durchaus verschiedene Definitionen verwendet werden. Während in zahlreichen, vor allem auch früheren Arbeiten zum Deutschen wie z.B. derjenigen von Weiß (1968: 55) ausschließlich der Punkt, das Frage- und das Ausrufezeichen als satzabschließende Interpunktionszeichen gewertet wurden, finden sich in Arbeiten zum Deutschen wie z.B. bei Niehaus (1997: 221) differenziertere Kriterien als Grundlage einer Satzdefinition: „Als satzabschließend gelten die folgenden Interpunktionszeichen: der Punkt, das Fragezeichen, das Ausrufezeichen. Der Doppelpunkt nimmt eine Sonderstellung ein, denn er wird nur dann als satzabschließendes Zeichen gewertet, wenn das erste Graphem des folgenden Wortes groß geschrieben wird". Wie zu sehen ist, wird zur Definition des Satzes bzw. des Satzendes jeweils ein anderes Inventar an Interpunktionszeichen herangezogen und zum Teil an bestimmte Kontextbedingungen gekoppelt. Als eine Schlussfolgerung daraus ergibt sich, dass solche Definitionsunterschiede nicht ohne Auswirkung auf quantitative Berechnungen bleiben. Zunächst aber gilt es, bevor wir derartige Definitionen automatisch auf die zu analysierenden Texte übertragen, die z.T. in der kulturspezifischen Tradition

[3] Vor der Diskussion der relevanten satzabschließenden Zeichen ist kurz auf die Rolle der Interpunktion in schriftlichen Texten und auf allfällige Editionsänderungen einzugehen. Als problematisch zu bezeichnen ist, dass die Interpunktion aufgrund von Editionseingriffen nicht immer der Autorenintention (vgl. Wake 1957: 334) entspricht. Wie etwa Janson (1964) in einer akribischen Studie zeigt, können unterschiedliche Editionen in der Setzung der Interpunktionszeichen stark divergieren. Dies impliziert jedoch auch, dass die durchschnittliche Satzlänge aufgrund unterschiedlicher Editionen in einem beträchtlichen Maß divergiert und somit keine verbindlichen Aussagen über die Satzlänge eines Autors getätigt werden können (vgl. Janson 1964: 29f.).

der Typographie begründete Spezifik in der Verwendung und Setzung von Interpunktionszeichen in den Texten der erwähnten Sprache aufzuzeigen.

Nach den Regeln der russischen Orthographie kommt dem Punkt die zentrale Funktion zu, das Ende von Sätzen zu markieren. Das Frage- und Ausrufezeichen dient, wie in anderen Sprachen auch, zur Kennzeichnung von Frage- und Ausrufesätzen sowie von Interjektionen. Unter Berücksichtigung der Wichtigkeit der erwähnten Interpunktionszeichen bei der Textgliederung könnte somit eine erste pragmatische Arbeitsdefinition wie folgt lauten:

Satzdefinition 1:

Ein Satz ist eine durch Punkt, Frage- oder Ausrufezeichen – ‚.', ‚?', ‚!' – abgegrenzte Einheit des Textes.

Es ist offensichtlich, dass eine solche Definition nur funktionieren kann und nur dann überhaupt sinnvoll ist, wenn die Datenbasis vor oder während der automatischen Analyse der Texte ‚manipuliert' wird, d.h. unter Berücksichtigung der Tatsache, dass der Punkt in der Funktion als Kennzeichnung von Abkürzungen (Beispiele: *т.е. = то есть*, *и т.д. = и так далее*, *и т.п. = и тому подобное*, u.a.m.,) und in der Form von Aufzählungen (Beispiel: *Но вы видите … до другого раза!*) nicht als in satzabschließender Position vorkommend gezählt wird. Während die Problematik der Abkürzungen sich gegebenenfalls durch eine automatisierte Auflösung in Form eines Abkürzungsverzeichnisses lösen lässt, bliebe die Schwierigkeit des Umgangs mit durch mehrere Punkte gekennzeichneten Aufzählungen bestehen. Noch problematischer an der Definition 1 wäre jedoch, dass Punkt, Fragezeichen und Ausrufezeichen nicht in allen Fällen unbedingt das Ende eines Satzes kennzeichnen müssen, so dass eine alternative Zählung sinnvoll scheint.

In Anlehnung an die einschlägigen Überlegungen von Grinbaum (1996) zur Automatisierung von Satzlängenuntersuchungen bietet es sich an, den Großbuchstaben als weiteres satzabgrenzendes Zeichen in eine formal bestimmbare Satzdefinition einzubauen. Wenn auch der Großbuchstabe im Russischen u.a. zur Kennzeichnung von Eigennamen, geographischen Bezeichnungen und Ähnlichem dient, liegt eine weitere zentrale Funktion des Großbuchstabens darin, den Anfang von Texten, Absätzen und einzelnen Sätzen zu markieren. In dieser Funktion als Gliederungsmerkmal von Texten können Großbuchstaben bei der Bestimmung von Satzgrenzen herangezogen werden (vgl. Grinbaum 1996: 454). Somit sind die Interpunktionszeichen [.], […], [?] und [!] in Kombination mit einem Großbuchstaben am Anfang des nächstfolgenden Satzes eindeutig als satzabschließend zu identifizieren. Da jedoch den erwähnten Interpunktionszeichen nicht in allen Fällen ein Buchstabe folgen muss (z.B. am Absatz- oder Textende), gelangt man zu der folgenden alternativen operationalen Satzdefinition:

Satzdefinition 2:

> *Ein Satz ist eine durch Punkt, Frage- oder Ausrufezeichen abgegrenzte Einheit des Textes, wobei [.], […], [?] und [!] als Satzendezeichen gelten, es sein denn, ein Kleinbuchstabe ist das erste Graphem des jeweils folgenden Wortes.*

Ohne Frage könnten weitere operationale Satzdefinitionen erarbeitet werden. Beschränkt man sich jedoch einstweilen auf die beiden obigen Definitionen, so ist es ganz offensichtlich, dass mit den beiden obigen der Satz und damit auch die Satzlänge recht zuverlässig automatisiert bestimmt werden kann. Dabei ist bei der Anwendung von Satzdefinition 2 zu erwarten, dass sich die absolute Anzahl der Sätze gegenüber Satzdefinition 1 verringert, so dass die unter dieser Bedingung verbleibenden Sätze im Vergleich zur Definition 1 durchschnittlich länger sind. Der Grund dafür ist darin zu sehen, dass auch Ausrufe- und Fragesätze innerhalb eines Satzgefüges als vollwertige Sätze betrachtet werden, wie das folgende Beispiel zeigt:

(Ex. 1): *«Ах, какая досада!», сказал Долгоруков.*

Hier würde Satzdefinition 2 es nach sich ziehen, die gesamte Sequenz als *einen* Satz zu behandeln, und nicht die der Autoren- bzw. Personenrede folgende fremde Rede als selbständigen Satz auszugliedern. Entsprechend würden nach Satzdefinition 1 zwei Sätze mit drei und zwei Wörtern gezählt, während auf der Basis von Satzdefinition 2 ein einziger Satz mit 5 Wörtern ausgewertet würde.

Natürlich könnte man bei der Definition von Satz weitere differenzierende Marker einführen und z.B. zusätzliche Restriktionen unter Einbeziehung von Anführungszeichen oder Tirets einführen. So ist es in der russischen Editionspraxis ja durchaus üblich, direkte Rede, die im Deutschen durch einfache oder doppelte, einleitende und ausleitende Anführungsstriche markiert wird, durch einleitende und ausleitende Tirets zu kennzeichnen:

(Ex. 2): *– Ах, какая досада! – сказал Долгоруков.*

Allerdings scheint in diesem Fall eine Orientierung an Satzdefinition 2 durchaus ausreichend, um Sequenzen wie die obige als einen einzigen vollständigen Satz zu identifizieren, da zwar die Aufeinanderfolge von Autoren- bzw. Personenrede und fremder Rede variabel ist – d.h. sowohl präponiert, postponiert als auch interponiert erfolgen kann –, und dass bei präponierter Autoren- bzw. Personenrede die Subjekt-Prädikat-Reihenfolge beibehalten wird, dass aber bei postponierter und interponierter Stellung die Aufeinanderfolge der Hauptsatzglieder invertiert wird, so dass in diesem Fall das Prädikat dem

Subjekt vorausgeht (so dass Eigennamen mit Großschreibung nicht als Markierung eines Satzanfangs fehlinterpretiert werden können).

Wie in anderen Fällen auch, ist es letztlich eine Frage der linguistischen Entscheidung bzw. der Konvention, eine der obigen Definitionen zu übernehmen oder zu modifizieren. Im Rahmen der vorliegenden Untersuchung wird vor dem Hintergrund der obigen Darlegungen und Begründungen eine Untersuchung der Satzlänge auf der Basis der Satzdefinition 2 erfolgen.

2.2. Satz: Maßeinheit

Auch die Wahl einer geeigneten Maßeinheit, die zur Bestimmung der Satzlänge herangezogen werden kann, ist de facto eine linguistische Entscheidung und damit nicht von der jeweils gewählten linguistischen Konzeption unabhängig. Im Hinblick auf die Wahl einer geeigneten Maßeinheit der Satzlänge erscheint die Messung in der Anzahl der Teilsätze bzw. der involvierten Phrasen pro Satz durchaus plausibel. Dies entspräche auch dem Postulat der Quantitativen Linguistik, demgemäß die Länge sprachlicher Konstrukte vorzugsweise in der Anzahl bzw. Länge ihrer unmittelbaren Konstituenten zu messen ist. Allerdings ist insbesondere für computerbasierte (automatisierte) Analysen eine nicht nur linguistisch begründbare, sondern auch technisch umsetzbare Definition des Teilsatzes (engl.: clause) notwendig.

Im Hinblick auf das Deutsche ist es in dieser Hinsicht durchaus üblich, sich – Bemerkungen von Winter (1964) folgend – auf das Verbvorkommen zu konzentrieren und die Anzahl finiter Verbformen – ausgedrückt durch die grammatischen Kategorien eines Verbs, insbesondere Person, Numerus, Tempus, Genus verbi und Modus – als Grundlage der Definition eines Teilsatzes heranzuziehen (vgl. Pieper 1979, Niehaus 1997); allerdings ist dieses Vorgehen aus linguistischer Sicht nicht unproblematisch, wie Grimm (1991) ausführlich gezeigt hat. Die entsprechende Diskussion muss hier nicht im Detail aufgerollt werden; jedoch wäre ein solches Vorgehen auch und gerade für das Russische aus linguistischer Sicht nicht problemlos, weil in diesem Fall nicht nur die im Russischen typischen Haupt- und Nebensätze ohne (Hilfs-)Verben unberücksichtigt blieben, sondern auch die insbesondere in schriftlichen Texten häufigen (vor allem post-positionalen) Partizipialkonstruktionen als Grundlage von Teil- bzw. Nebensätzen.

Vor diesem Hintergrund ist es verständlich, dass die meisten bisherigen Arbeiten zur Satzlänge im Russischen incl. der eingangs erwähnten auf der Maßeinheit ‚Wort pro Satz' beruhen, auch wenn dies im Grunde genommen dem Überspringen einer linguistischen Ebene gleichkommt.

Lediglich Roukk (2001a) hat in ihrer Arbeit zur Satzlänge in Erzählungen Čechovs versucht, die Vorkommenshäufigkeit von Satzlängen auf der Basis

von Teilsätzen pro Satz zu bestimmen, in Anlehnung an die oben dargestellten Definitionen einen Teilsatz definierend als „eine syntaktische Konstruktion, die eine finite Verbform enthält oder – bei Ellipsen – ergänzen lässt“ (Roukk 2001a: 114).

Die von Roukk (2008) vorgenommene Übertragung und Anwendung dieser Teilsatzdefinition auf die Untersuchung der Längenverhältnisse – also auf die Beziehung zwischen der (in der Anzahl der Teilsätze pro Satz gemessenen) Satzlänge und der (in der Anzahl der Wörter pro Satz gemessenen) Teilsatzlänge – hat allerdings eindeutige Negativ-Befunde gezeitigt. Roukk (2007: 609) hat deswegen geschlussfolgert, dass das Menzerath-Altmann-Gesetz für das Russische nicht gültig sei, wobei sie selbst zwei alternative Erklärungen dieses Befunds anbietet, die beide sprachspezifisch ausgerichtet sind: 1. das Gesetz funktioniere prinzipiell nicht, zumindest nicht in der gegenwärtigen Form, für das Russische, 2. für das Russische sei eine besondere Zählweise (d.h. Bestimmung der Maßeinheit) der Konstituenten (d.h. in diesem Fall: der Teilsätze) nötig. Beide Annahmen schließen einander freilich nicht aus. Hinzu kommt noch der Umstand, dass der Untersuchung von Roukk nur das 17. Kapitel aus dem IV. Buch von Tolstojs *Anna Karenina* als russisches originalsprachiges Untersuchungsmaterial zugrunde lag. Dieses umfasst nach der Zählung von Roukk insgesamt 231 Sätze – eine Datenbasis, die als nicht gerade breit anzusehen ist, wobei eine re-analysierende Zählung gemäß der obigen Satzdefinition 2 sogar nur 199 Sätze ergibt.

Abgesehen davon, dass es somit die Untersuchung dieser Frage auf jeden Fall auf eine breitere Materialbasis zu stellen gilt, scheint vor dem Hintergrund der obigen Diskussion Vieles für die zweite Alternative zu sprechen, dass nämlich die Definition der Teilsätze einer nicht unwesentlichen Korrektur bedarf: Bleiben nämlich die Partizipialkonstruktionen unberücksichtigt, und spielen diese von ihrer Vorkommenshäufigkeit her eine nicht unwesentliche Rolle, so ergibt sich zwangsläufig eine Schieflage im Datenmaterial, die durchaus für die beobachteten Negativbefunde verantwortlich sein könnte. Diese Annahme erhält zusätzliche Plausibilität vor dem Hintergrund der Befunde von Buk/Rovenchak (2008) zur Abhängigkeit der Teilsatzlänge von der Satzlänge in ukrainischen Texten, die aufgrund einer modifizierten Teilsatzbestimmung überzeugende empirische Anhaltspunkte für die Wirksamkeit des Menzerath-Altmann-Gesetzes im Ukrainischen liefern.

Der Nachweis, dass dies in gleicher Weise auch das Russische betrifft, fehlt allerdings nach wie vor. Aufgrund der bislang fehlenden technischen Voraussetzung, Teilsätze und deren Längen mit den von Buk/Rovenchak für das Ukrainische verwendeten Definitionen auch am Russischen auf automatisierter Ebene an umfangreichem Datenmaterial zu analysieren, bietet es sich – zumindest im Sinne einer Approximation an die Problematik – im hier gegebenen Kontext aus pragmatischen bzw. technischen Gründen an, eine

durchaus plausible Alternative zu verfolgen, nämlich die Satzlänge in der Anzahl der Wörter pro Satz zu messen.

Dieses Verfahren steht, wie einleitend gesagt wurde, in der Tradition der Satzlängenforschung zum Russischen allgemein und wurde u.a. auch von Roukk (2001b) verwendet. Abgesehen von der (technisch) leichteren Praktikabilität lässt es sich durchaus auch theoretisch rechtfertigen: Erstens sind diesem Fall keine zusätzlichen syntaktischen Analysen notwendig – deren Ergebnisse ihrerseits jeweils theorieabhängig divergieren würden –, und zweitens führen erfahrungsgemäß beide Vorgangsweisen – also Messung in der Anzahl der Teilsätze ebenso wie in der Anzahl der Wörter pro Satz – zum Nachweis analoger sprachlicher Regularitäten. Freilich ist bei der Bezugnahme auf die Wortebene als Maßeinheit für Satzlänge eine deutlich höhere Varianz zu erwarten als bei Teilsätzen, so dass möglicherweise spezifische Verfahren der Datenzusammenfassung anzuwenden sein werden, um den Daten immanente Strukturen von der Tendenz her besser erkennbar darzustellen. Insofern wird in der vorliegenden Arbeit davon ausgegangen, dass die Anzahl der Wörter pro Satz eine für die gegebenen Fragestellung adäquate Maßeinheit ist, ein Vorgehen, das seinerseits allerdings eine angemessene Definition dessen, was unter ‚Wort' zu verstehen ist, erfordert.

2.3. Wort: Definition

Auch im Hinblick auf eine Definition dessen, was allgemein unter ‚Wort' zu verstehen ist (Wurzel 2000), kann man nicht von einer eindeutigen a priori-Definition ausgehen. In computergestützten Textanalysen mit automatisierten Vorgangsweisen ist es üblich, sich auf eine orthographische Definition des Wortes zu stützen. Demgemäß ist ein Wort folglich als eine durch Satzzeichen bzw. durch eine Leerstelle abgegrenzte Einheit des Textes zu verstehen. Eine solche graphematische Wortdefinition ist freilich nicht nur für das Deutsche, wie Fuhrhop (2008) anschaulich zeigt, problematisch. Auch und gerade für das Russische (und analog auch für andere slawische Sprachen) ergibt sich vielmehr das Problem, dass man es als Folge einer derartigen Definition z.B. bei Präpositionen wie ‚с', ‚к' ‚в' u.a. mit einer eigenen Klasse von nullsilbigen Wörtern zu tun hätte, wenn man das Vorhandensein eines Vokals (bzw. eines Diphthongs) oder eines silbischen Liquids (bzw. Konsonanten in silbenbildender Funktion) als Basis einer Silbe ansieht – zur Problematik der nullsilbigen Wörter vgl. Antić, Kelih, Grzybek (2006).

Alternativ böte sich vor diesem Hintergrund eine phonologische Wortdefinition an, der gemäß unter einem als einer Taktgruppe verstandenen Wort eine aus einer akzentogenen Wortform und gegebenenfalls einem oder auch mehr als einem Pro- bzw. Enklitikon bestehende (Text-)Einheit zu verstehen

ist (vgl. Lehfeldt 1999). Eine solche Definition wäre auch und gerade für das Russische (und andere slawische Sprachen) insofern plausibel, als unter dieser Bedingung u.a. auch Akzentverschiebungen auf (graphematischen Wörtern entsprechende) Wortformen einheitlich erfasst werden.

Allerdings sind Analysen auf der Basis einer derartigen Definition technisch nicht ohne weiteres automatisiert durchzuführen, sondern erfordern eine umfangreiche Analyse der Akzentstruktur des untersuchten Materials. Insofern hat sich in den vergangenen Jahren zumindest im Hinblick auf slawische Sprachen eine orthographisch-phonetische Wortdefinition als effizient erwiesen, bei welcher die genannten nullsilbigen Wörter als Klitika behandelt und dem jeweils folgenden Lexem zugeordnet werden. Die Auswirkungen auf die Anzahl der unterschiedenen Einheiten und deren Länge in Abhängigkeit von der Wahl einer der drei genannten Wortdefinitionen lässt sich an folgendem Beispiel veranschaulichen – vgl. Kelih (2007), der die Auswirkung der Wortdefinition auf die sprachlichen Verhältnisse im Detail dargelegt hat:

(Ex.3a) │Он│ │вслушивался│ │в│ │трубку│,│волнуясь│ │и│ │стараясь│ │уловить│ │что│ │то│ │такое,│что│ │могли│ бы│ │скрывать│ │от│ │него│.

(Ex.3b) │Он│ │вслушивался│ │втрубку│,│волнуясь│ │и│ стараясь│ │уловить│ │что│ │то│ │такое,│что│ могли│ │бы│ │скрывать│ │от│ │него│.

(Ex. 3c) │Он│ │вслушивался│ │втрубку,│ │волнуясь│ │истараясь│ │уловить│ │чтото│ │такое│,│что│ │моглибы│ │скрывать│ │отнего│.

Es ist offensichtlich, in welchem Maße sich die Wahl einer der Wortdefinitionen zunächst auf die ergebende Wortlänge und damit natürlich auch auf die Wortlängenverteilung auswirkt, in weiterer Folge dann aber natürlich auch auf die Untersuchung des Zusammenhangs zwischen Wort- und Satzlänge. Wenn der vorliegenden Untersuchung die dritte, also die orthographisch-phonetische Wortdefinition zugrunde gelegt werden soll, verbleibt im Rahmen der operationalen Definitionen noch eine Bestimmung einer geeigneten Maßeinheit der Wortlänge.[4]

[4] Natürlich gibt es in diesem Zusammenhang eine Reihe weiterer Entscheidungen zu treffen, u.a. zum Beispiel, wie mit Zahlen („между 12 и 2 часами“; „Георгий 1-й степени“), mit einer nicht auflösbaren Abkürzung wie „N.N.“, mit fremdsprachigen Einschüben oder Sequenzen („Maman не хочет этого“) umzugehen ist u.v.a.m., was hier nicht im Einzelnen diskutiert werden kann und muss.

2.4. Wortlänge: Maßeinheiten

Im Hinblick auf die Wahl einer geeigneten Maßeinheit bei der Bestimmung der Wortlänge ist es im Bereich der Informationstechnologie nach wie vor durchaus üblich, die Länge eines Wortes in der Anzahl der Buchstaben pro Wort zu berechnen. Ein solches Vorgehen ist aus linguistischer Sicht fragwürdig, unabhängig davon ob man Buchstaben oder Grapheme zugrunde legt, ob man dabei Digraphen oder Trigraphen differenziert oder nicht; es entspricht auch nicht dem Postulat der Quantitativen Linguistik, ein Konstrukt (möglichst) in der Länge seiner Konstituenten zu messen, weshalb auch eine Berechnung in der Anzahl der Phoneme (oder Phone) pro Wort nicht als zielführend anzusehen ist, abgesehen davon, dass dazu Prozeduren wie etwa Graphem-Phonem-Konvertierungen nötig wären. Alternativ ließe sich die Wortlänge in der Anzahl der Silben oder der Morpheme pro Wort berechnen. Während für die Berechnung in der Anzahl der Morpheme allerdings eine (wiederum stark theorieabhängige) morphologische Segmentierung des sprachlichen Analysematerials notwendig wäre, lässt sich die Silbenanzahl pro Wort relativ einfach über die Anzahl der Vokale pro Wort berechnen, was zumindest im Russischen ohne weitere Zusatzprozeduren möglich ist, da es weder Diphthonge noch silbische Liquide (bzw. Konsonanten in silbenbildender Funktion) gibt. Aus diesem Grunde soll im vorliegenden Text Wortlänge in der Anzahl der Silben pro Wort berechnet werden.

2.5. Definitionen und Maßeinheiten: Zusammenfassung

Wie sich aus den obigen Darstellungen (2.1.–2.5) ableiten lässt, haben wir damit die für die quantitative Analyse und Modellbildung notwendigen operationalen Definitionen geleistet. Ausgehend von der Satzdefinition 2 – der zufolge ein Satz eine durch Punkt, Frage- oder Ausrufezeichen und folgenden Großbuchstaben abgeschlossene Einheit des Textes ist – wird die Satzlänge in der Anzahl der auf orthographisch-phonetischer Ebene bestimmten Wörter bestimmt, deren Länge in der Anzahl der Silben pro Wort berechnet wird.

Damit können wir zur eigentlichen Fragestellung nach einem systematischen Zusammenhang zwischen Satz- und Wortlänge im Russischen zurückkehren. Vor der Untersuchung dieses Zusammenhangs scheint es jedoch geboten, zunächst jeweils die Vorkommenshäufigkeiten von Wortlängen sowie von Satzlängen jeweils auf eigene Regularitäten hin zu untersuchen.

3. Vorkommenshäufigkeit von Satzlägen und Wortlängen: Empirische Befunde und theoretische Modellierung

3.1. Satzlängenhäufigkeit

Auf der Basis der obigen Satzdefinition gelangt man bei der Analyse der Satzlängen von Tolstojs *Анна Каренина* zu dem Resultat, dass dieser Text insgesamt 19297 Sätze aufweist, von denen der kürzeste aus genau einem Wort besteht, der längste aus 151 Wörtern. Die durchschnittliche Satzlänge beträgt $\bar{x} = 13.89$ Wörter pro Satz bei einer Standardabweichung von $s = 11.08$. Tab. 1 stellt die Vorkommenshäufigkeiten (f_x) der jeweils aus x Wörtern bestehenden Sätze zusammenfassend dar.

Tab. 1: Vorkommenshäufigkeiten von Satzlängen in *Анна Каренина*

x	f_x	x	f_x	x	f_x	x	f_x	x	f_x
1	340	21	320	41	49	61	14	84	1
2	725	22	330	42	50	62	4	87	1
3	1093	23	271	43	39	63	9	88	1
4	1296	24	244	44	41	64	5	90	1
5	1294	25	230	45	30	65	5	91	1
6	1145	26	199	46	39	66	4	92	2
7	1115	27	194	47	27	67	3	98	1
8	992	28	189	48	26	68	7	99	1
9	975	29	159	49	24	69	4	100	1
10	874	30	145	50	26	70	4	106	1
11	840	31	130	51	21	71	6	116	1
12	736	32	115	52	17	72	1	125	1
13	648	33	106	53	25	73	2	138	1
14	643	34	91	54	15	74	4	151	1
15	598	35	81	55	11	75	1		
16	534	36	86	56	12	77	3		
17	489	37	78	57	11	78	3		
18	489	38	72	58	12	79	2		
19	373	39	55	59	7	80	1		
20	362	40	53	60	8	81	1		

Abb. 1 veranschaulicht die Häufigkeitsverteilung in graphischer Form – auf der *x*-Achse sind die Satzlängen abgetragen – der besseren Anschaulichkeit halber ist die Darstellung der Satzlänge auf den Bereich der bis 100 Wörter langen Sätze beschränkt –, auf der *y*-Achse die jeweiligen Vorkommenshäufigkeiten.

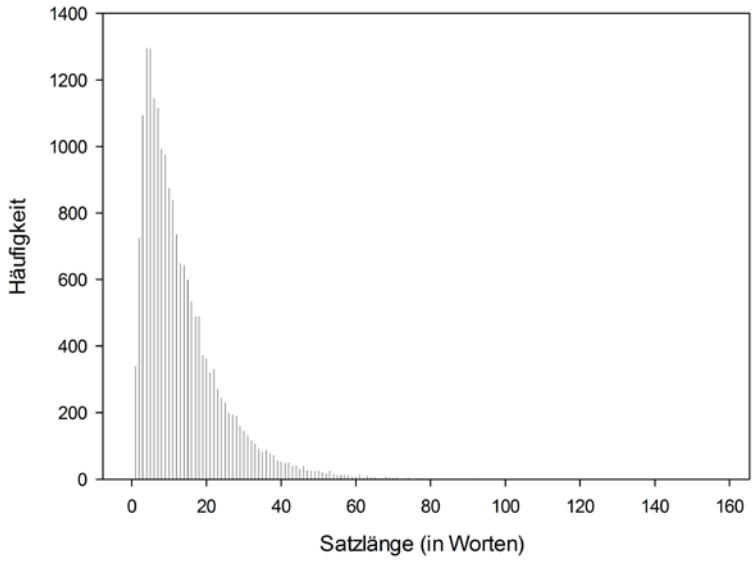
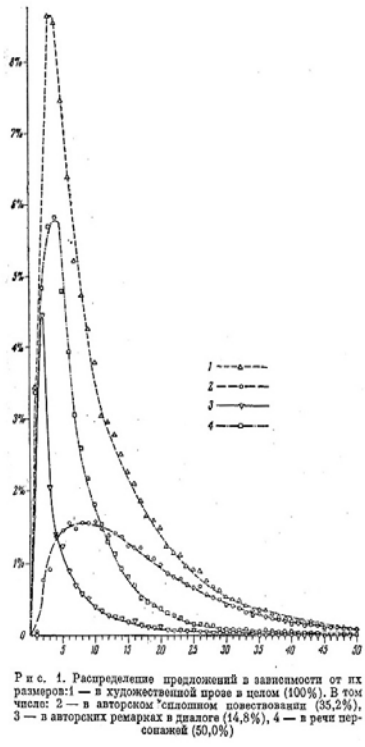


Abb. 1: Häufigkeitsverteilung der Satzlängen in Tolstojs *Анна Каренина*

Deutlich zu sehen ist die typische Rechtsschiefe (Linkssteilheit) der Häufigkeitsverteilung: Es gibt insgesamt deutlich mehr kürzere als längere Sätze. Für das Russische hat Leskiss bereits in den 60er Jahren versucht, ein theoretisches Modell zu finden, das sich im Wesentlichen an der symmetrischen Normalverteilung orientiert und logarithmische Transformationen beinhaltet.
Abb. 2 veranschaulicht am Beispiel künstlerischer Texte die von Leskiss herausgearbeiteten Satzlängenhäufigkeiten, sowohl für den Gesamttext als auch differenziert nach narrativen Sequenzen, auktorialen Einschüben und dialogischer Figurenrede.



Р и с. 1. Распределение предложений в зависимости от их размеров: 1 — в художественной прозе в целом (100%). В том числе: 2 — в авторском сплошном повествовании (35,2%), 3 — в авторских ремарках в диалоге (14,8%), 4 — в речи персонажей (50,0%)

Abb. 2: Häufigkeitsverteilung von Satzlängen im Russischen nach Leskiss (1962)

Abgesehen davon, dass Leskiss dieses Modell (wie durchaus noch üblich zu seiner Zeit) nicht durch entsprechende statistische Tests auf seine Eignung geprüft hat, würde man heute nicht mehr so vorgehen und die Schiefe der Verteilung nicht als Abweichung von einer Normalverteilung interpretieren, sondern als für sprachliche Daten genuin charakteristisch ansehen und in der Folge ohne transformierte Variablen zu modellieren versuchen.

Roukk (2001b) hat – ausgehend von obiger Satzdefinition 1 und einer orthographischen Wortdefinition – an ausgewählten literarischen, journalistischen und wissenschaftlichen Texten herausgearbeitet, dass die Hyperpoisson-Verteilung

$$(1) \qquad P_x = \frac{a^x}{b^{(x)}{}_1F_1(1;b;a)}$$

ein geeignetes Modell ist, Satzlängenhäufigkeiten im Russischen zu modellieren, und zwar – da es keine Sätze mit 0 Wörtern gibt – in ihrer 1-verschobenen Form:

$$(2) \qquad P_x = \frac{a^{x-1}}{b^{(x-1)}{}_1F_1(1;b;a)}$$

Im Vergleich dazu hat Kelih (2001) an umfangreicherem russischen Textmaterial – unter Verwendung von obiger Satzdefinition 2 und auf der Basis einer orthographischen-phonetischen Wortdefinition – zeigen können, dass die negative Binomialverteilung

$$(3) \qquad P_x = \binom{k+x-1}{x} p^k q^x \qquad \text{x = 0, 1, 2, …}$$

ein geeignetes Modell ist Satzlängenhäufigkeiten des Russischen theoretisch zu beschreiben, ebenfalls natürlich in 1-verschobener Form:

$$(4) \qquad P_x = \binom{k+x-2}{x-1} p^k q^{x-1} \qquad x = 1, 2, 3, \ldots$$

wobei dieses Modell aufgrund der jeweils veränderlichen Parameterwerte flexibel genug ist, auch Satzlängenhäufigkeiten verschiedener Autoren oder Textsorten zu modellieren. In unserem konkreten Fall von Tolstojs *Анна Каренина* bietet es sich allerdings an, dieses Modell in einer etwas komplexeren Form zu verwenden, die sich als Mischung zweier Verteilungen ergibt, die beide jeweils einer negativen Binomialverteilung folgen, allerdings jede von ihnen mit unterschiedlichen Parameterwerten für *k* und *p* (und *q = 1–p*), so dass sich eine (1-verschobene) gemischte negative Binomialverteilung mit den Gewichtungen α und 1– α ergibt:

$$(5) \qquad P_x = \alpha \binom{k_1+x-2}{x-1} p_1^{k_1} q_1^{x-1} + (1-\alpha) \binom{k_2+x-2}{x-1} p_2^{k_2} q_2^{x-1} \qquad x = 1, 2, 3, \ldots$$

Der Umstand, dass ein solches Mischmodell sich im Vergleich zum „einfachen“ Modell als effizienter erweist, ist rechnerisch darin zu sehen, dass ein Modell mit mehr (und damit auch mit mehr zu interpretierenden) Parametern natürlich in der Regel zu besseren Anpassungsergebnissen führt. Inhaltlich dürfte dies damit zu begründen sein, dass der Romantext heterogene Satzre-

gimes mit jeweils eigenen Längenverteilungen aufweist, z.B. dialogische vs. narrative oder deskriptive Sequenzen – dazu gibt es bislang allerdings keinerlei systematische Untersuchungen.

Abb. 3 stellt das Ergebnis[5] in graphischer Form dar, das sich mit den Parameterwerten $k = 2.47$, $p_1 = 0.26$, $p_2 = 0.12$ und dem Gewichtungsfaktor $\alpha = 0.52$ ergibt.

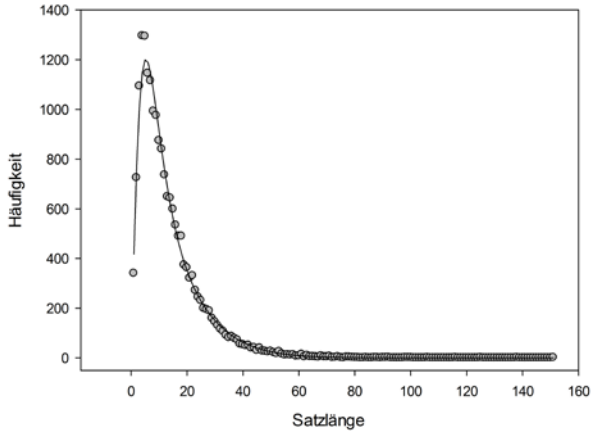


Abb. 3: Satzlängenhäufigkeiten in L.N. Tolstojs *Анна Каренина.* Anpassung der gemischten negativen Binomialverteilung

Die Anpassung dieses Verteilungsmodells an die Daten stellt sich als ausgezeichnet dar, was sich rechnerisch aus dem Wert des Diskrepanzkoeffizienten $C = X^2 / N$ ablesen lässt, der als Indikator für die Güte der Anpassung dient, und der im gegebenen Fall mit $C = 0.0075$ als sehr gut anzusehen ist.[6]

Damit kann die erste unserer als Voraussetzung anzusehenden Forderung als nachgewiesen gelten, nämlich dass die Verteilung der Satzlängen in *Анна Каренина* nicht chaotisch ist, sondern einer bekannten Regularität folgt und gesetzmäßig organisiert ist.

---

[5] Um die Anzahl der Parameter zu reduzieren, wird die gemischte negative Binomialverteilung im vorliegenden Fall mit $k = k_1 = k_2$ berechnet.

[6] Für einen Wert von $C < 0.02$ geht man von einer guten, für $C < 0.01$ von einer sehr guten Anpassung aus.

3.2. Wortlängenhäufigkeit

Auch zur Wortlängenhäufigkeit erübrigt es sich, an dieser Stelle viele Worte zu verlieren, da die entsprechende Problematik in den letzten Jahren ausführlich diskutiert und aufgearbeitet worden ist (vgl. Grzybek 2006). Somit können wir uns sogleich der Analyse von Tolstojs *Анна Каренина* zuwenden.

Unter Berücksichtigung der obigen Definitionen kommen im gesamten Text 258384 Wörter vor. Deren Länge beträgt im Durchschnitt $\bar{x} = 2.22$ Silben pro Wort bei einer Standardabweichung von $s = 1.18$. Tab. 2 führt die Vorkommenshäufigkeit (f_x) für die einzelnen Wortlängen (x) an, die sich im Intervall von $x_{min} = 1$ bis $x_{max} = 10$ Silben pro Wort erstrecken.

Tab. 2 stellt die Vorkommenshäufigkeiten (f_x) der jeweils aus x Silben bestehenden Wörter zusammenfassend dar – die in der dritten Spalte angeführten, mit Np_x bezeichneten Werte werden weiter unten erläutert und können hier einstweilen ignoriert werden.

Tab. 2: Vorkommenshäufigkeiten von Wortlängen in *Анна Каренина*

x	f_x	Np_x
1	84629	86960
2	85649	82134
3	50241	51717
4	25338	24423
5	9828	9227
6	2257	2905
7	370	784
8	65	185
9	6	39
10	1	7

Abb. 4 veranschaulicht die Häufigkeitsverteilung in graphischer Form:

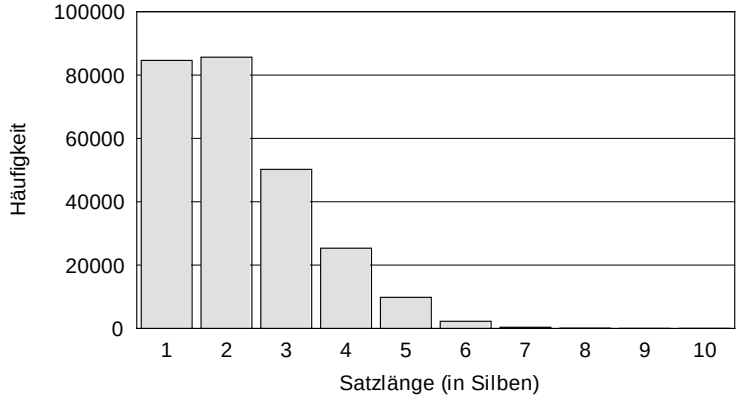


Abb. 4: Wortlängenhäufigkeiten in L.N. Tolstojs *Anna Karenina*

Es stellt sich in weiterer Folge die Frage, inwiefern die beobachteten Häufigkeiten der Wortlängen einem theoretischen Modell folgen. Auch bezüglich dieser Frage gibt es im Hinblick auf die verschiedenen slawischen Sprachen, u.a. auch für das Russische, eine ganze Reihe von Arbeiten, in denen überwiegend verschiedene Modifikationen oder aber Verallgemeinerungen der (1-verschobenen) Poisson-Verteilung

$$P_x = \frac{e^{-a}a^{x-1}}{(x-1)!} \qquad x = 1, 2, 3, \ldots \tag{6}$$

angewendet wurden. Dabei handelt es sich zum einen um sog. „lokale" Modifikationen wie z.B. die Singh-Poisson-Verteilung (s. Wimmer/Altmann 1999: 605f.)

$$P_x = \begin{cases} 1-\alpha+\alpha e^{-a} \\ \dfrac{\alpha a^{x-1}e^{-a}}{(x-1)} \end{cases} \qquad x = 1, 2, 3, \ldots \tag{7}$$

bei welcher die erste Klasse eigens modelliert und die übrigen Klassen entsprechend angepasst werden, oder um Verallgemeinerungen wie z.B. die Hyperpoisson-Verteilung (s. Wimmer/Altmann 1999: 281f.)

$$P_x = \frac{a^{x-1}}{{}_1F_1\left(1;b;a\right)} \qquad x = 1, 2, 3, \ldots \tag{8}$$

die einen weiteren Parameter b aufweist. Dieses Vorgehen ist durchaus üblich, und beide Modelle sind in der Quantitativen Linguistik mehrfach zur Modellierung von Wortlängen- und anderen Häufigkeiten verwendet worden (vgl. Best 2001); es erfordert jedoch in Ergänzung zur Poisson-Verteilung als *dem* Modell für vom Zufall abhängige Ereignisse die linguistische Begründung der notwendigen Modifikation der ersten Klasse bzw. des zusätzlich zu interpretierenden Parameters.

Im Falle unserer Daten von Tolstojs *Анна Каренина* ist jedoch die einparametrige Poisson-Verteilung ausreichend[7], allerdings in ihrer links-gestutzten Form[8], die auch als positive Poisson-Verteilung bezeichnet wird:

$$P_x = \frac{e^{-a}a^x}{x!\left(1-e^{-a}\right)} \qquad x = 1, 2, 3, \ldots \tag{9}$$

Mit einem Parameterwert von $a = 1.89$ ergeben sich die in der 3. Spalte der Tab. 2 (s.o.) dargestellten theoretischen Werte (Np_x), die in Abb. 5 graphisch dargestellt sind; die grau unterlegten Balken stehen für die beobachteten Daten (f_x), die weißen für die theoretischen (Np_x).

[7] Freilich lassen sich mit mehr-parametrigen Modellen auch hier noch bessere Anpassungsergebnisse erzielen; der Unterschied ist jedoch im gegebenen Fall nur geringfügig, weshalb dem Modell mit den wenigsten Parametern der Vorzug zu geben ist.

[8] Während bei einer Verschiebung um eine Position das gesamte Verteilungsmodell um 1 nach rechts verschoben wird, basiert eine Links-Stutzung auf der Annahme, dass die Besetzung der Klasse $x = 0$ praktisch nicht möglich ist, so dass diese Klasse bei der theoretischen Modellierung „eliminiert“ wird.

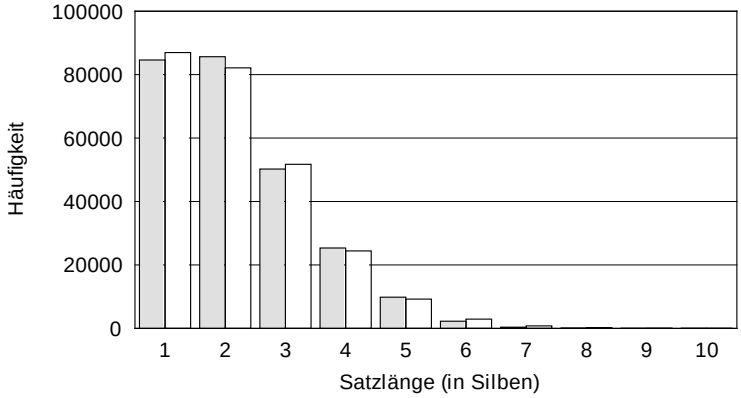


Abb. 5: Wortlängenhäufigkeiten in L.N. Tolstojs *Anna Karenina*: Anpassung der positiven Poisson-Verteilung

Wie aus der Abb. 5 ersichtlich ist, stimmen die beobachteten Werte (f_x) und die sich aus (5) ergebenden theoretischen Werte Np_x gut überein, was durch den Werte des Diskrepanzkoeffizienten $C = 0.003$ (vgl. hierzu Fußnote 5) rechnerisch bestätigt wird.

Damit kann auch die zweite unserer als Voraussetzung anzusehenden Forderungen als nachgewiesen gelten, nämlich dass die Verteilung der Wortlängen in *Анна Каренина* nicht chaotisch ist, sondern einer bekannten Regularität folgt und gesetzmäßig organisiert ist.

4. Abhängigkeit zwischen Satzlänge und Wortlänge

Damit können wir uns nach der notwendigen Klärung aller Voraussetzungen nunmehr der Frage nach dem Zusammenhang von Satz- und Wortlänge (gemäß den oben vorgenommenen Definitionen) zuwenden.

4.1. Satzlänge ↔ Wortlänge in *Анна Каренина*

Tab. 3 zeigt die Ergebnisse der entsprechenden Berechnungen in jeweils drei zusammengehörigen Spalten: in der ersten (*SaL*) ist die jeweilige Satzlänge angegeben, in der zweiten deren Vorkommenshäufigkeiten (*f*), die sich auch schon in Tab. 1 (s.o.) finden, und in Ergänzung dazu die durchschnittliche Wortlänge (*WoL*) für die jeweilige Satzlänge.

Tab. 3: Zusammenhang von Satzlängen und Wortlängen in *Анна Каренина*

SaL	*f*	*WoL*	*SaL*	*f*	*WoL*	*SaL*	*f*	*WoL*	*SaL*	*f*	*WoL*
1	340	2.2647	26	199	2.2578	51	21	2.3081	77	3	2.3853
2	725	2.1359	27	194	2.2518	52	17	2.3032	78	3	2.1880
3	1093	2.0494	28	189	2.2132	53	25	2.2936	79	2	2.4810
4	1296	2.0355	29	159	2.2720	54	15	2.1432	80	1	2.4500
5	1294	2.1037	30	145	2.2614	55	11	2.3306	81	1	2.1975
6	1145	2.1007	31	130	2.2759	56	12	2.2188	84	1	2.3810
7	1115	2.1363	32	115	2.2508	57	11	2.1722	87	1	2.3448
8	992	2.1569	33	106	2.2573	58	12	2.2716	88	1	1.9318
9	975	2.1850	34	91	2.3284	59	7	2.3850	90	1	2.2444
10	874	2.1684	35	81	2.2790	60	8	2.2500	91	1	2.4286
11	840	2.1871	36	86	2.2474	61	14	2.2951	92	2	2.1957
12	736	2.2183	37	78	2.2367	62	4	2.2540	98	1	2.6429
13	648	2.2204	38	72	2.2376	63	9	2.3086	99	1	2.5051
14	643	2.2253	39	55	2.1939	64	5	2.3844	100	1	2.2700
15	598	2.2317	40	53	2.2552	65	5	2.3723	106	1	2.1792
16	534	2.2328	41	49	2.2339	66	4	2.1174	116	1	2.1724
17	489	2.2283	42	50	2.3248	67	3	2.4030	125	1	2.3200
18	489	2.2295	43	39	2.2302	68	7	2.2647	138	1	2.8768
19	373	2.2585	44	41	2.2955	69	4	2.1884	151	1	2.9470
20	362	2.2297	45	30	2.3237	70	4	2.3071			
21	320	2.2579	46	39	2.2737	71	6	2.1174			
22	330	2.2625	47	27	2.1875	72	1	2.1528			
23	271	2.2784	48	26	2.2131	73	2	2.1233			
24	244	2.2609	49	24	2.2645	74	4	2.4527			
25	230	2.2610	50	26	2.1677	75	1	2.2267			

Abb. 6 stellt die Ergebnisse in anschaulicher Form dar.

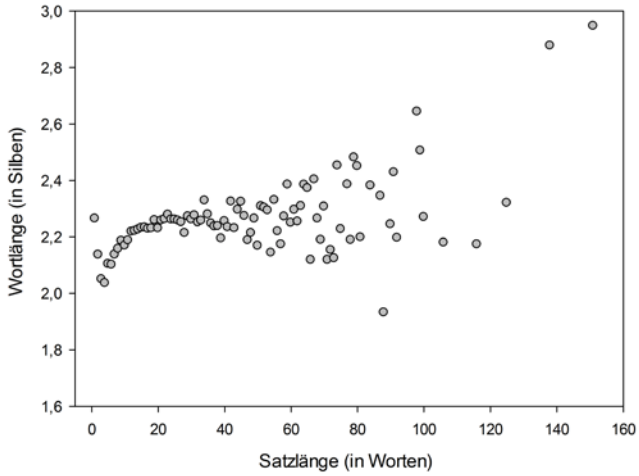


Abb. 6: Zusammenhang von Satzlängen und Wortlängen in *Анна Каренина*

Eine Reihe von Beobachtungen lassen sich auf den ersten Blick machen:

a. Es ist offensichtlich, dass ab einer Satzlänge von ca. 33 Wörtern pro Satz die Streuung der durchschnittlichen Wortlänge enorm zunimmt, so dass eine allgemeine Tendenz ab dieser Satzlänge nicht ohne weiteres erkenntlich ist bzw. erkennbar bleibt.
b. Die besonders kurzen Sätze (von 1 bis ca. 3-4 Wörter) zeichnen offenbar durch eine zur zentralen Gesamttrend gegenläufige Tendenz aus, insofern diese eine durchschnittlich größere und bis zur Satzlänge von ca. 3-4 Wörtern pro Satz abnehmende Wortlänge aufweisen.
c. Im zentralen Bereich der Satzlängen von ca. 3-4 Wörtern pro Satz bis zu ca. 30 Wörter pro Satz liegt eine klare Tendenz zunehmender Wortlänge mit ansteigender Satzlänge vor.

Alle drei Beobachtungen sind im Prinzip nicht unbekannt: In ihren Untersuchungen zur Abhängigkeit von Satz- und Wortlänge in verschiedenen russischen Textsorten – 44 dramatischen Texten, 120 Privatbriefen, 69 literarischen Texten und 60 Zeitungskommentaren, sowie einem aus diesen Texten kompilierten, aus insgesamt 23539 Sätzen bestehenden Korpus – haben Grzybek et al. (2008) zum Teil ähnliche Beobachtungen gemacht.

Ad a: Im Hinblick auf die Satzlänge, ab welcher die Streuung deutlich zunimmt, sprechen Grzybek et al. (2008) von einem „upper critical point“ (*UCL*); dieser divergiert allerdings über die verschiedenen Textsorten im Bereich von *UCL*=22 für die Dramen bis zu *UCL*=32 für die Kommentare; für das gesamte Korpus lag er bei *UCL*=40. Zur Erklärung dieses Phänomens verweisen Grzybek et al. (2008: 116) auf zwei verschiedene Optionen, eine statistische (i) und eine linguistische (ii):

i. Der statistischen Erklärung zufolge wäre annehmbar, dass die Schwankung auf eine unzureichende Anzahl an Beobachtungen pro Datenpunkt zurückgeht, die der Bildung des Mittelwerts der Wortlänge für eine gegebene Satzlänge zugrunde liegt; Grzybek et al. (2008) haben auf der Grundlage ihrer Befunde im Hinblick auf diese Annahme vermutet, dass empirisch eine Häufigkeit von f_{SL}> 30 notwendig sein könnte, damit es zu stabilen Ergebnissen kommt. Die Ergebnisse der vorliegenden Analyse zu Anna Karenina bekräftigen diese Erklärungsoption allerdings nicht, denn wie der Tab. 3 zu entnehmen ist, ist diese Bedingung bis zur Satzlänge von *SL* = 46 erfüllt, bis zur Satzlänge von *SL*=33 basiert der jeweilige Mittelwert sogar auf mindestens $f_{SL} \geq 100$.

ii. Die linguistische Erklärungsoption beinhaltet zunächst den Umstand der Berechnung von Satzlänge in der Anzahl der Wörter pro Satz, was de facto der Überspringung einer Ebene (nämlich des Teilsatzes) gleichkommt; in Anbetracht dieser Tatsache ist es sogar eher erstaunlich, dass bis zur Satzlänge von ca. *SL*=33 eine relativ stabile Tendenz auszumachen zu sein scheint. Auch und gerade vor diesem Hintergrund wäre es aus linguistischer Perspektive begründbar, dass die Mechanismen der Selbstregulation ab einer bestimmten Satzlänge nicht mehr greifen, da sie sich der steuernden (mitnichten intentionalen) Kontrolle des Produzenten entziehen. Grzybek et al. (2008) verweisen in diesem Zusammenhang auf Limits der menschlichen Informationsverarbeitung und Begrenzungen der Gedächtnisspanne, was zurückführt zu Millers (1956) „magischer Zahl“ von 7±2 und deren linguistisch-syntaktischer Interpretation von Yngve (1960, 1996), und durchaus im Einklang mit neuesten Überlegungen zur Quantitativen Syntax (Köhler 1999, 2007, 2012) steht. Geht man nämlich – bei allem Vorbehalt – mit Roukk (2007) davon aus, dass Teilsätze im Russischen durchschnittlich ca. 4-5 Wörter aufweisen, dann ist bei komplexen Sätzen mit sieben Teilsätzen ziemlich genau der oben

empirisch festgestellte $UCL \approx 30$ erreicht. Sollte diese Annahme zutreffen, müsste dies in weiterer Folge hypothetisch dazu führen, dass sich die durchschnittliche Wortlänge ab diesem *UCL* zufällig um einen relativ konstanten Wert herum streut, wobei das Ausmaß der Streuung von der jeweiligen Anzahl der beobachteten Datenpunkte abhängt.

Grzybek et al. (2008) haben in ihrer ersten Annäherung an die Frage einer theoretischen Modellierung der Daten zur Erfassung der „core data structure" die Erfüllung von zwei Bedingungen vorausgesetzt, nämlich (a) das Vorhandensein von $f_{SL} > 30$ sowie (b) eine Satzlänge von $SL \leq 30$. Den obigen Resultaten Rechnung tragend, scheint es geboten, die erste der beiden Bedingungen aufzugeben; sinnvoll hingegen scheint es, Datenpunkte mit $SL > 30$ einstweilen aus der Betrachtung auszuschließen, da deren Tendenz ganz offensichtlich nicht ohne spezifische Verfahren der Datenglättung erfasst werden kann.

Ad b: Im Hinblick auf die kurzen Sätze sprechen Grzybek et al. (2008) von einem „lower critical point" (LCP), bis zu welchem die Wortlänge bei zunehmender Satzlänge zunächst abnimmt; dieser liegt für die verschiedenen untersuchten Textsorten im Bereich von $2 \leq LCP \leq 7$, für das gesamte Korpus bei $LCP = 4$. Auch im Falle von *Anna Karenina* erweist sich $LCP = 4$ als zutreffend.

Grzybek et al. (2008) haben in ihrer ersten Annäherung an die Frage einer theoretischen Modellierung der Daten die entsprechenden kurzen Sätze aus ihrer Betrachtung ausgeschlossen, um sich so auf das zu konzentrieren, was sie als „core data structure" bezeichnen. Allerdings scheint die unter (c) angeführte Beobachtung aus quantitativ-linguistischer bzw. informationstheoretischer Sicht nicht überraschend: Kurze Sätze mit wenig Wörtern müssen, wenn sie nicht – wie z.B. Repliken à la „да" oder „нет" –, von der Frequenz her redundant sind, im Rahmen des minimalen zur Verfügung stehenden Umfangs viel Information enthalten, was bekanntermaßen durch die Verwendung seltener bzw. längerer Wörter realisiert wird. Da Kurzsätze zudem im Bereich der Syntax in der Regel keine hyper- bzw. hypotaktischen Gliederungen aufweisen, enthalten sie relativ wenige Synsemantika, die ihrerseits durch spezifische Kürze charakterisiert sind. Insofern ist ihre gegenläufige Tendenz linguistisch plausibel. Sinnvoll wäre es jedoch, wenn beide Tendenzen, also sowohl die anfangs absteigende als auch die im zentralen Bereich ansteigende, in toto modelliert werden könnten, ohne dass die Kurzsätze aus dem Modellierungsansatz ausgeklammert oder aber „autonom" modelliert werden müssten.

Ad c: Der für *Anna Karenina* zu beobachtende Anstieg der Wortlänge mit zunehmender Satzlänge im zentralen Bereich konnte ebenfalls schon von Grzybek et al. (2008) beobachtet werden. Dabei traten allerdings z.T. überraschende textsortenspezifische Unterschiede zutage (die im hier vorliegenden Text nicht weiter berücksichtigt werden können): Wie sich nämlich herausstellte, galt die ansteigende Tendenz nur für die literarischen Prosatexte, nicht hingegen für die journalistischen und dramatischen Texte und die Privatbriefe, bei denen die Wortlänge sich im zentralen Bereich relativ stabil als mehr oder weniger konstant herausstellte.

Damit ergibt sich im Gesamtverlauf unserer Studie als nächste Aufgabe, ein theoretisches Modell zu suchen, welches den Zusammenhang zwischen Wort- und Satzlänge im Bereich von $1 \leq SL \leq 33$ abdeckt (wobei die Begrenzung auf den *UCP* = 33 sich im gegebenen Fall mit der Frequenz von $f_{SL} > 100$ stützen lässt).

4.2. Wort- und Satzlänge im Lichte des Menzerath-Altmann-Gesetzes

In seiner allgemeinsten Form ist das, was heute als Menzerath-Altmann bekannt ist, von Altmann (1980) in seinen „Prolegomena to Menzerath's Law" vorgeschlagen worden. Dieses basiert zunächst auf der Annahme einer Proportionalitätsrelation zwischen Konstrukt und Konstituenten, genauer: der konstanten Abnahme der Länge der Konstituenten bei zunehmender Länge des Konstrukts. Drückt man diese Annahme der Abnahme mathematisch als $y' = -a$ aus, resultiert dies in der Differentialgleichung

$$\frac{y'}{y} = -a \tag{10}$$

mit der Lösung

$$y = Ke^{-ax}. \tag{11a}$$

Um auch komplexere Fälle mit einem anfänglichen Anstieg bis zu einem Maximum bei $x \neq 0$ erfassen zu können, hat Altmann (1980) eine Erweiterung der Differentialgleichung (10) durch die Hinzufügung einer inversen Proportionalitätskomponente vorgeschlagen, so dass sich dann aufgrund der Differentialgleichung

$$\frac{y'}{y} = -a + \frac{b}{x} \tag{12}$$

für $b = 0$ die obige Lösung (11a) ergibt, der zufolge die Länge von Komponenten eine Funktion der Länge sprachlicher Konstrukte ist. Für $b \neq 0$ gibt es zwei weitere Optionen, nämlich für $a = 0$

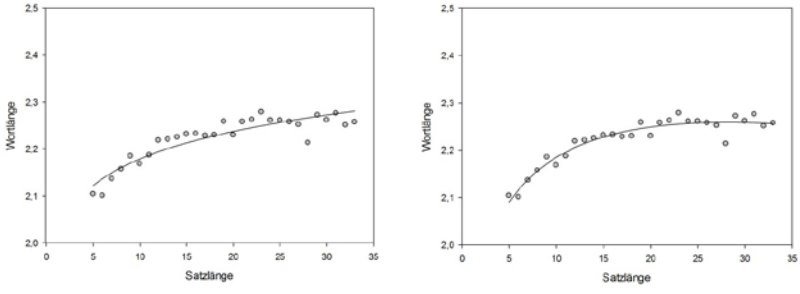
(11b) $$y = Kx^b ,$$

und für $a \neq 0$

(11c) $$y = Kx^b e^{-ax} .$$

Für linguistische Zwecke ist oft (11b) gewissermaßen als „Standardform" des Menzerath-Altmann-Gesetzes angesehen worden: Mit $b < 0$, sagt sie eine Abnahme der Länge (oder auch Komplexität) linguistischer Konstituenten bei einer Zunahme der Länge bzw. Komplexität des durch sie gebildeten Konstrukts voraus.

Würde man sich, wie Grzybek et al. (2008), auf die Modellierung der „core data structure" ($4 \leq SL \leq 30$) von *Анна Каренина* beschränken, käme man mit diesem Modell (11b) und den Parameterwerten $a = 2.00$ und $b = 0.0381$auf eine Anpassung, die mit einem Determinationskoeffizienten von $R^2 = 0.84$ als einigermaßen geeignet anzusehen wäre (vgl. Abb. 7a). Im Vergleich dazu würde mit (11c), das einen Parameter mehr hat, ein bestens geeignetes Modell vorliegen, welches bei Parameterwerten von $C = 1.85$, $b = 0.0858$ und $a = 0.0031$ einen auf eine ausgezeichnete Anpassung hinweisenden Determinationskoeffizienten von $R^2 = 0.92$ aufweist (vgl. Abb. 7b).



a) Für $4 \leq SL \leq 30$: $y = ax^b$

b) Für $1 \leq SL \leq 30$: $y = C \cdot x^{a1} \cdot exp(-a0 \cdot x)$

Abb. 7: Modellierung des Zusammenhangs von Satz- und Wortlänge in *Анна Каренина* (I)

Allerdings ist keines dieser Modelle geeignet, die Datenstruktur im gesamten Bereich $1 \leq SL \leq 30$ abzudecken; hierzu ist offenbar ein komplexeres Modell notwendig. Erst in jüngster Zeit haben Wimmer und Altmann (2005, 2006) den obigen Ansatz in ihrer „Allgemeinen Ableitung einiger sprachlicher Gesetze“ integriert und damit Erweiterungen und Verallgemeinerungen eingeführt, die bislang nur in Teilbereichen zur Anwendung gekommen sind. Der Ansatz beruht auf der Differentialgleichung

$$\frac{y'}{y} = \left(a + \frac{b}{x} + \frac{c}{x^2} + \frac{d}{x^3} + ... \right). \tag{13}$$

Wie zu sehen ist, ergibt sich die obige Differentialgleichung (12) für $c,d,\ldots = 0$ aus (13), diese hat ansonsten für $a,b,c,\ldots \neq 0$ die Lösung

$$y = Ke^{ax} x^b e^{-c/x - d/2x^2 - \ldots}. \tag{14}$$

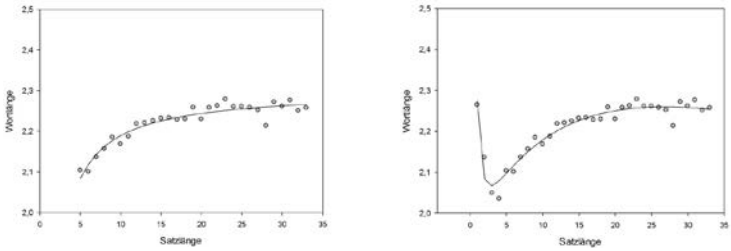
Entsprechend ergeben sich auch aus (14) die Gleichungen (11a–c). Hinzu kommt mit $e^{d/x}$ jedoch ein weiterer (optionaler) Faktor, der sich aus (14) mit $d = 0$ und ergibt. In der Zusammenschau liegt somit ein System von insgesamt sechs Funktionen mit maximal vier Parametern (K, a,b,c) vor, mit denen auch komplexere sprachliche Verhältnisse modelliert werden können. Die komplexeste von ihnen ist (IV), alle anderen lassen sich als Spezialfälle von (IV) für bestimmte Parameterwerte bzw. -konstellationen verstehen. Tab. 4 führt für jede der sechs Gleichungen die Parameterkonstellation und die sich daraus jeweils ergebende Gesamtanzahl der Parameter an.

Tab. 4: Die Funktionen des Menzerath-Altmann-Gesetzes und dessen Erweiterungen

I	$y = K \cdot e^{ax}$	$a < 0$, $b,c = 0$	2
II	$y = K \cdot x^b$	$b < 0$, $a,c = 0$	2
III	$y = K \cdot e^{ax} \cdot x^b$	$a,b \neq 0$, $c = 0$	3
IV	$y = K \cdot e^{-c/x}$	$c > 0$	2
V	$y = K \cdot x^b \cdot e^{c/x}$	$b,c \neq 0$	3
VI	$y = K \cdot e^{ax-c/x} \cdot x^b$	$a,b,c \neq 0$	4

Interessanterweise kommt man mit dem zwei-parametrigen Modell (IV) für die „core data structure“ mit den Parameterwerten $K = 2.30$ und $c = 0.49$ auf ein vergleichbar gutes Ergebnis von $R^2 = 0.92$ wie mit dem in Abb. 8a dargestellten Modell (III) bzw. (11c), obwohl letzteres einen Parameter mehr aufweist.

Zur zusätzlichen und gleichzeitigen Integration der Kurzsätze in ein einheitliches Modell ist allerdings die Bezugnahme auf (VI) nötig, mit welchem ein ausgezeichnetes Modell vorliegt, das mit Parameterwerten $K = 1.74$, $a = 0.0038$, $b = 0.1098$ und $c = 0.1526$ auf einen Wert von $R^2 = 0.92$ kommt (vgl. Abb. 8b).



a) Für $4 \leq SL \leq 30$: $y = K \cdot e^{c/x}$ **b)** Für $1 \leq SL \leq 30$: $y = K \cdot e^{ax} \cdot x^b \cdot e^{-c/x}$

Abb. 8: Modellierung des Zusammenhangs von Satz- und Wortlänge in *Анна Каренина* (II)

Damit kann – erstmals – für das Russische die Beziehung zwischen Satz- und Wortlänge als im Rahmen des Altmann-Menzerath-Gesetzes als nachgewiesen gelten. Die allfällige Annahme einer diesbezüglichen Sonderstellung des Russischen kann damit als obsolet betrachtet werden. Allerdings gilt, diese wesentliche Schlussfolgerung in einer Reihe von Punkten zu spezifizieren.

Erstens ist die Gültigkeit aus (quantitativ-)linguistischer Sicht unter der Überspringung einer Ebene, nämlich der Ebene des Teilsatzes (bzw. clause, phrase, o.a.m.) nachgewiesen; in diesem Sinne lässt sich vom Menzerath-Altmann-Gesetz 2. Ordnung sprechen. Dies ist insofern von Bedeutung, als vor dem Hintergrund des Menzerath-Altmann-Gesetzes die beobachtete Satzlänge-Wortlänge-Relation plausibel ist: bei zunehmender Satzlänge werden die Teilsätze kürzer, und kürzere Teilsätze haben die Tendenz, aus längeren Wörtern gebildet zu werden, so dass in der Zusammenschau die Wortlänge mit der Satzlänge ansteigt – allerdings sieht die mathematische Formulierung des Menzerath-Altmann-Gesetzes als eines stochastischen Gesetzes keine Transitivität vor (vgl. Altmann/Meyer 2005). Insofern können die obigen Befunde, streng genommen, kein zwingender Nachweis dafür sein, dass das

Menzerath-Altmann-Gesetz im Russischen auch auf der Satzlänge-Teilsatzlänge wirksam ist – ein starkes Argument dafür ist es freilich dennoch, und zugleich eine Bestärkung der Annahme, dass die von Roukk (2007) dargestellten Beobachtungen eher durch unpassende Teilsatzdefinitionen bedingt und/oder durch zu geringes Datenmaterial begründbar sind.

Die auf den ersten Blick erstaunlichen Ergebnisse, dass (a) zur Modellierung der „core data structure" eine zweiparametrige Funktion der erweiterten Menzerath'schen Funktionen (Tab. 4, IV) ebenso geeignet ist wie eine dreiparametrige der „üblichen" (Tab. 4, III), und dass (b) zur Integration auch der Kurzsätze die Hinzufügung nicht nur eines, sondern zweier Parameter nötig ist, erhält eine plausible Erklärung, wenn man sich die obigen Differentialgleichungen (12) und (13) und die Erläuterungen dazu noch einmal anschaut. Die in Abb. 8a dargestellte Funktion für die „core data structure" beruht auf der Differentialgleichung

$$\text{(13a)} \qquad \frac{y'}{y} = a + \frac{c}{x^2}\,.$$

In diesem Fall ist offensichtlich keine „Korrektur" für die gegenläufige „Störung" der Kurzsätze erforderlich (weil b=0 und damit b/x nicht in die Funktion eingeht); wohl aber ist offenbar in diesem Fall in Ergänzung zu der einfachen Konstanten $-a$ die sich aus $c \neq 0$ ergebende quadratische Komponente c/x^2 aus (13) nötig, und es liegt nahe, sie als durch die Überspringung der Teilsatzebenen bedingten „Störfaktor" zu interpretieren. Dies würde auch erklären, warum zur Erfassung der Gesamtstruktur (vgl. Abb. 8b) sowohl $b,c \neq 0$ und die sich damit ergebende Differentialgleichung

$$\text{(13b)} \qquad \frac{y'}{y} = a + \frac{b}{x} + \frac{c}{x^2}$$

nötig ist – weil nämlich sowohl die durch die Kurzsätze bedingte gegenläufige „Störungstendenz" als auch der Ebenensprung erfasst werden müssen. Damit läge eine ebenso plausible wie vollständige Interpretation des gesamten Modells vor, die es freilich an umfangreicherem Material unter Berücksichtigung allfällig modifizierender Faktoren zu überprüfen und in weiterer Folge auch an anderen Sprachen zu testen gilt.

5. Resümee und Ausblick

Die oben dargestellten Analysen widmen sich im Wesentlichen der Frage nach einem systematischen Zusammenhang zwischen Satz- und Wortlänge.

Diese Frage verortet sich in einem breiteren Horizont im Kontext der Problematik (selbst-regulierender) Wechselbeziehungen zwischen sprachlichen Einheiten bzw. Konstrukten verschiedener Ebenen, wie sie vor allem in der Quantitativen und/oder Synergetischen Linguistik reflektiert werden. Von besonderer Bedeutung sind in diesem Zusammenhang vor allem die Annahmen des bekannten Menzerath-Altmann-Gesetzes, denen zufolge die ein sprachliches Konstrukt konstituierenden Einheiten umso kürzer sind, je länger das betreffende Konstrukt ist; dabei gilt es allerdings zu beachten, dass diese Annahme sich auf die jeweils *direkten* Konstituenten eines Konstrukts beziehen, was im Falle der Satzebene Kategorien wie Phrasen oder Teilsätze wären, die damit gleichzeitig als Maßeinheit für die Satzlänge anzusehen wären. Allerdings wurde in den letzten Jahren aufgrund von Untersuchungsergebnissen zu Satz-Teilsatz-Relationen im Russischen die Vermutung geäußert, dass dieses Gesetz womöglich auf das Russische nicht zutreffe.

Im vorliegenden Text, einem Beitrag zur Beleuchtung und Klärung dieser Frage, wird eingangs aufgezeigt, dass den berichteten Beobachtungen und den damit verknüpften Schlussfolgerungen verschiedene Faktoren zugrunde liegen können, die nicht unbedingt sprachspezifische Charakteristika beinhalten, sondern gegebenenfalls auch mit unzureichendem Datenmaterial oder unpassenden Definitionen auf der Wort- und/oder Teilsatzebene ihre Erklärung finden könnten. In diesem Fall sollte die Untersuchung der Längenbeziehungen zwischen Satz und Wort (also unter Überspringung der direkten Konstitutionsebene) Anhaltspunkte im Hinblick auf eine Lösung der Problematik bereitstellen und im positiven Fall – also bei Vorliegen regulärer Satz-Wort-Beziehungen – zumindest indirekte Argumente für die Wirksamkeit des Menzerath-Altmann-Gesetzes auch im Russischen liefern.

Am Beispiel von L.N. Tolstojs Roman *Анна Каренина* ist deshalb Gegenstand der vorliegenden Untersuchung die Frage, ob solche Zusammenhänge zwischen der Satz- und Wortlänge bestehen und, wenn ja, ob diese sich darüber hinaus ebenfalls im Rahmen des bekannten Menzerath-Altmann-Gesetzes interpretieren lassen. In Anbetracht der erhaltenen Befunde lässt sich im Hinblick auf diese Frage eine Reihe von Feststellungen treffen bzw. von Schlussfolgerungen ziehen:

1. Bei der notwendigerweise zu klärenden Voraussetzung, dass die Häufigkeiten sowohl der Satzlängen als auch der Wortlängen jeweils eigenen Regularitäten folgen und nicht chaotisch organisiert sind, konnte gezeigt werden, dass beide Voraussetzungen gegeben sind und jeweils Modellen folgen, die aus der bisherigen einschlägigen Forschung bestens bekannt sind:

a. Die Häufigkeiten der Wortlängen folgen der (positiven) Poisson-Verteilung.
b. Für die Satzlänge erweist sich das Modell der (gemischten) negativen Binomial-Verteilung als geeignet.

2. Das Verhältnis von Satz- und Wortlänge ist – zumindest bis zu einer Satzlänge von ca. $SL \leq 30$ Wörter pro Satz – nicht chaotisch, sondern regulär organisiert:

   a. Die Länge von Wörtern in Kurzsätzen im Intervall von $1 \leq SL \leq 4$ hat die Tendenz, mit zunehmender Satzlänge abzunehmen.
   b. Die Länge von Wörtern in Sätzen im Intervall von $5 \leq SL \leq 30$ hat die Tendenz, mit zunehmender Satzlänge zuzunehmen.
   c. Während es in den früheren Untersuchungen als möglich angesehen wurde, dass die hohen Streuungen im Bereich der Satzlänge $SL > 30$ mit der geringen Frequenz ($f_{SL} < 30$) von Vorkommnissen, welche die Grundlage der durchschnittlichen Satzlänge darstellen, zusammen hängen, erscheint es auf der wesentlich breiteren Materialbasis (mit $f_{SL} \geq 30$ bis $SL = 46$) der hiermit vorgelegten Ergebnisse plausibler, dass diese Streuung linguistisch motiviert ist: Für Satzlängen mit $SL > 30$ scheinen die Mechanismen der Selbstregulierung sich zunehmend zu verlieren und nicht mehr wirksam zu sein.

3. Die Regularitäten des Zusammenhangs von Satz- und Wortlänge lassen sich im Rahmen der von Wimmer/Altmann entwickelten „Unified Theory of Some Linguistic Laws“ modellieren, aus der sich im Vergleich zu den ursprünglichen Funktionen des Menzerath-Altmann-Gesetzes eine Reihe erweiterter Funktionen ableiten lassen. Dabei konnte im vorliegenden Text erstmals eine Reihe von Besonderheiten beobachtet werden:

   a. Frühere Studien zur Beziehung der Satz- und Wortlänge (im Russischen) haben sich bei der Modellierung unter Auslassung der Kurzsätze auf die „core data structure“ im Bereich von $5 \leq SL \leq 30$ beschränkt;
   b. dabei konnte der Zusammenhang zwischen Wort- und Satzlänge im Rahmen des Menzerath-Altmann-Gesetzes modelliert werden,
   c. Im vorliegenden Text gelingt es erstmals, die gesamte Datenstruktur (bis zur Satzlänge von $SL \leq 30$) in einem einheitlichen Modell zu erfassen.
   d. Zur Erfassung der gesamten Datenstruktur reichen die Funktionen des ursprünglichen Menzerath-Altmann-Gesetzes nicht aus; vielmehr muss auf erweiterte Funktionen Bezug genommen werden, die sich

aus der von Wimmer/Altmann entwickelten „Unifed Theory of Linguistic Laws“ ergeben.

e. Die Bezugnahme auf diese erweiterten Funktionen erlaubt auch eine modifizierte Erfassung der „core data structure“ im Rahmen eines Modells mit weniger Parametern, das im Sinne Occam’s vorzuziehen ist und bessere Interpretationsoptionen bereit stellt.

Mit den durchgeführten Untersuchungen und dem vorgestellten Modell lässt es sich als nachgewiesen betrachten, dass die Satzlänge (auch im Russischen) keine isolierte Größe ist, sondern reguläre Beziehungen zu Einheiten und Konstrukten anderer Ebenen aufweist. Die Tatsache, dass dies im vorliegenden Fall im Hinblick auf die Wortlänge nachgewiesen werden konnte, ist zwar kein zwingender Beweis, dass entsprechende Beziehungen auch zwischen Satz- und Teilsatzlänge vorliegen, aber ein überaus starkes Argument für diese Annahme. Dies wiederum würde dafür sprechen, dass in der letzten Zeit vorgebrachte Vermutungen, das Menzerath-Altmann-Gesetz treffe auf das Russische nicht zu, nicht haltbar sind, sondern in anderen statistischen und/oder linguistischen Problemen begründet sind – einer unzureichenden Materialbasis und/oder unzutreffenden (linguistischen) Definitionen.

Literatur

Altmann, Gabriel (1980): “Prolegomena to Menzerath’s Law.” In: *Glottometrika 2*. Bochum, 1–10.

Altmann, Gabriel (1983): „H. Arens’ «Verborgene Ordnung» und das Menzerathsche Gesetz.“ In: Faust, Manfred et al. (Hrsg.), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik:* Tübingen: Narr; 31–39.

Altmann, Gabriel (1988): „Verteilungen der Satzlängen.“ In: Schulz, Klaus P. (ed.), *Glottometrika 9*. Bochum: Brockmeyer; 147–161.

Altmann, Gabriel; Lehfeldt, Werner (1980): *Quantitative Phonologie*. Bochum: Brockmeyer.

Altmann, Gabriel; Meyer, Peter (2005): “Physicists look at language.” In: Altmann, Gabriel; Levickij, Viktor; Perebyjnis, Valentina (eds.), *Проблеми квантитативної лингвістики*. *Problems of Quantitative Linguistics*. Černivci: Ruta; 42–59.

Altmann, Gabriel; Schwibbe, Michael H. (1989): *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms.

Antić, Gordana; Kelih, Emmerich; Grzybek, Peter (2006): “Zero-syllable Words in Determining Word Length.” In: Grzybek, Peter (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht, NL: Springer; 117–156.

Best, Karl-Heinz (ed.) (2001): *Häufigkeitsverteilungen in Texten.* Göttingen: Peust & Gutschmidt.

Buk, Solomija; Rovenchak, Andrij (2008): "Menzerath-Altmann Law for Syntactic Structures in Ukrainian", in: *Glottotheory, 1*; 10–17.

Bünting, Klaus Dieter; Bergenholtz, Henning (1995): *Einführung in die Syntax.* Stuttgart: Beltz.

Cramer, Irene M. (2005): „Das Menzerathsche Gesetz." In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Raimund G. (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein Internationales Handbuch – An International Handbook*. Berlin, New York: de Gruyter; 650–688.

Fuhrhop, Nanna (2008): „Das graphematische Wort (im Deutschen): Eine erste Annäherung", in: *Zeitschrift für Sprachwissenschaft, 27/2*; 189–228

Grimm, Christian (1991): *Zum Mythos Individualstil: mikrostilistische Untersuchungen zu Thomas Mann*. Würzburg: Königshausen & Neumann.

Grzybek, Peter; Kelih, Emmerich; Stadlober, Ernst (2008): "The relation between word length and sentence length: an intra-systemic perspective in the core data structure", in: *Glottometrics, 16;* 111–121.

Grzybek, Peter; Stadlober, Ernst; Kelih, Emmerich (2007): "The Relationship of Word Length and Sentence Length. The Inter-Textual Perspective." In: Decker, Reinhold; Lenz, Hans-J. (eds.), *Advances in Data Analysis*. Berlin, Heidelberg: Springer; 611-618.

Grzybek, Peter (2012a): "Michail Lopatto (1892–1981)." in: *Glottometrics*. [In prep.]

Grzybek, Peter (2012b): „Der Satz und seine Beziehungen II: Satzlänge und Kapitellänge im Russischen (Am Beispiel von L.N. Tolstojs «Анна Каренина»)." [In prep.]

Hřebíček, Luděk (1995): *Text levels. Langauge Constructs, Constituents, and the Menzerath-Altmann Law*. Trier: wvt.

Kelih, Emmerich (2002): *Untersuchungen zur Satzlänge in russischen und slowenischen Prosatexten. Band 1 & Band 2*. M.A. Thesis, Graz.

Kelih, Emmerich (2007): „Zur Frage der Wortdefinitionen in Wortlängenuntersuchungen." In: Kaliuscenko, Volodymir; Köhler, Reinhard, Levickij, Viktor (eds.): *Problems of Typological and Quantitative Lexicology*. Chernivtsi: Ruta, 91–105.

Kelih, Emmerich; Grzybek, Peter (2005): „Satzlänge: Definitionen, Häufigkeiten, Modelle. (Am Beispiel slowenischer Prosatexte)." In: *Quantitative Methoden in Computerlinguistik und Sprachtechnologie.* [Special Issue of: *LDV-Forum. Zeitschrift für Computerlinguistik und Sprachtechnologie // Journal for Computational Linguistics and Language Technology, 20*; 31–51.

Köhler, Reinhard (1999): „Syntactic structures: properties and interrelations", in: *Journal of Quantitative Linguistics, 6*; 46–57.

Köhler, Reinhard (2007): „Quantitative analysis of syntactic structures.“ In: Mehler, Alexander; Köhler, Reinhard (eds.), *Aspects of Automatic Text Analysis*. Berlin, Heidelberg: Springer; 191–209.
Köhler, Reinhard (2012): *Quantitative Syntax Analysis*. Berlin, Boston: De Gruyter Mouton.
Lehfeldt, Werner (1999) : „Akzent.“ In Jachnow, Helmut (ed.), *Handbuch der sprachwissenschaftlichen Russistik und ihrer Grenzdisziplinen.* Wiesbaden: Harrassowitz; 34–48.
Leskiss, G.A. (1962): „O razmerach predloženija v russkoj naučnoj i chudožestvennoj proze 60-ch godov XIX v.“, in: *Voprosy jazykoznanija, 2*; 78–95.
Leskiss, G.A. (1963): „O zavisimosti meždu razmerom predloženija i charakterom teksta“, in: *Voprosy jazykoznanija, 3*; 92–112.
Leskiss, G.A. (1964): „O zavisimosti meždu razmerom predloženija i ego strukturoj v raznych vidach teksta“, in: *Voprosy jazykoznanija, 3*; 92–112.
Neumann, Susanne (2009): *Das Menzerath-Altmann-Gesetz als Textcharakteristikum.* M.A. Thesis, Trier.
Niehaus, B. (1997): „Untersuchung zur Satzlängenhäufigkeit.“ In: Best, Karl-Heinz (ed.), *Glottometrika 16*. Trier: wvt; 213–276.
Pieper, Ursula (1979): *Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse*. Tübingen: Narr.
Roukk, Maria (2001a): „Satzlängen in Texten von A. Tschechow.“ In: *Göttinger Beiträge zur Sprachwissenschaft, 5*; 113–120.
Roukk, Maria (2001b): „Satzlängen im Russischen“. In: Best, Karl-Heinz (ed.), *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt; 211–218.
Roukk, Maria (2008): “The Menzerath-Altmann Law in translated texts as compared to the original texts.” In: Grzybek, Peter; Köhler, Reinhard (eds.), *Exact Methods in the Study of Language and Text*. Berlin etc.: Mouton de Gruyter.
Sherman, Lucius A. (1888): “Some observations upon the sentence-length in English prose”, in: *University of Nebraska Studies, 1*; 119–130.
Weiß, H. (1968): *Statistische Untersuchungen über Satzlänge und Satzgliederung als autorenspezifische Stilmerkmale. (Beitrag zur mathematischen Analyse der Formalstruktur von Texten).* Doktorarbeit, TH Aachen.
Williams, C.B. (1940). “A note on the statistical analysis of sentence-length as a criterion of literary style“, in: *Biometrika, 31*; 356–361.
Wimmer, Gejza; Altmann, Gabriel (2005): “Unified derivation of some linguistic laws.” In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Raimund G. (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein*

Internationales Handbuch – An International Handbook. Berlin, New York: de Gruyter; 791–807.
Wimmer, Gejza; Altmann, Gabriel (2006): "Towards a unified derivation of some linguistic laws." In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues*. Dordrecht, NL: Springer; 329–337.
Winter, Werner (1964): "Styles as dialects". In: Lunt, Horace G. (ed.), *Proceedings of the 9th International Congress of Linguistics*. The Hague: Mouton; 324–330.
Yule, Udney G. (1938/39): "On sentence length as a statistical characteristic of style in prose: with application to two cases of disputed authorship", in*: Biometrika, 30*; 363–390.
Yngve, Victor H. (1960): "A model and a hypothesis for language structure", in: *Proceedings of the American Philosophical Society,* 194; 444–466.
Yngve, Victor H. (1960): *From Grammar to Science: New Foundations for General Linguistics*. Amsterdam, Philadelphia: Benjamins.

peter.grzybek@uni-graz.at