Emmerich Kelih, Peter Grzybek

Satzlänge: Definitionen, Häufigkeiten, Modelle (Am Beispiel slowenischer Prosatexte)

1 Einleitung

Die vorliegende Untersuchung versteht sich als ein Beitrag zur Satzlängenforschung. Nach einleitender Darstellung der Analysemöglichkeiten auf der Ebene der Satzlängen, geht es hauptsächlich um die Diskussion der Anwendung von unterschiedlichen Satzdefinitionen. Auf der Basis eines Korpus slowenischer Texte wird der Frage nachgegangen, welchen Einfluss die Anwendung unterschiedlicher (durchaus üblicher) Satzdefinitionen auf (a) deskriptive Kenngrößen der Häufigkeitsverteilung hat, und (b) inwiefern davon die Adäquatheit und Güte theoretischer Verteilungsmodelle abhängt.

2 Satzlänge: Stilcharakteristikum – Textklassifikation – Modellierung

Untersuchungen zu Satzlängen werden mit den unterschiedlichsten Fragestellungen im Rahmen eines breiten Forschungsspektrums durchgeführt: Zum Einen wird die Satzlänge in erster Linie als ein (1) spezifisches Stilcharakteristikum betrachtet; zum Anderen wird die Häufigkeitsverteilung von Satzlängen vor allem darauf hin untersucht, inwiefern sich diese durch (2) theoretische Wahrscheinlichkeitsmodelle beschreiben lässt. Beide Bereiche[1] sind im folgenden einleitend etwas ausführlicher zu kommentieren.

ad 1. Im Bereich von stilistischen Untersuchungen ist die Bedeutung der Satzlänge früh erkannt worden (vgl. Sherman, 1888). Ohne an dieser Stelle einen erschöpfenden Überblick geben zu wollen, lässt sich zeigen, dass sich in der aktuellen Forschung vor allem zwei miteinander verbundene Bereiche herauskristallisieren, in denen intensiv mit der Satzlänge gearbeitet wird. Es ist dies der Bereich der quantitativen Klassifikation von Texten (vgl. Mistrík, 1973; Pieper, 1979; Karlgren und Cutting, 1994; Bolton und Roberts, 1995). Gemeinsam ist diesen Untersuchungen, dass in der Regel versucht wird, aufgrund der (durchschnittlichen) Satzlänge und einer ganzen Reihe weiterer quantitativ erfassbarer Textmerkmale bestimmte Textsorten, Genre-Gruppen, Diskurs-Stile u.ä. zu identifizieren. Einen Spezialbereich innerhalb der Textklassifikation nimmt die Frage ein, inwiefern die Satzlänge als eine Teilgröße eines "linguistischen Fingerabdruckes" zu verstehen ist und somit als Merkmal bei der Bestimmung der Autorschaft herangezogen werden kann (vgl. Yule, 1939; Smith, 1983; Kjetsaa, 1984; Holmes, 1994).

[1] Nicht weiter vorgestellt werden soll die Diskussion der Satzlänge im Bereich von psycholinguistischen Untersuchungen. In diesen wird – ausgehend von der als Indikator der grammatikalisch-syntaktischen Komplexität verstandenen Satzlänge – versucht, selbige als Grad der "Lesbarkeit" bzw. "Textverständlichkeit" qualitativ zu interpretieren (vgl. dazu Teigeler, 1968; Flesch, 1948; Tuldava, 1993).

ad 2. Parallel zu den soeben genannten Fragestellungen rückt die Diskussion in den Vordergrund, inwiefern sich die Verteilung von Satzlängen in adäquater Weise durch theoretische Verteilungsmodelle beschrieben werden kann. Derartige[2] Arbeiten wurden bereits in den 40ern Jahre des 20. Jhd.s durchgeführt (vgl. Yule, 1939; Williams, 1940) und werden bis in die Gegenwart hinein diskutiert (so z.B. von Sichel (1974); Sigurd et al. (2004) u.a.). Einen grundlegenden Neuansatz in der Modellierung der Satzlängenverteilung liefert Altmann (1988b,a), der die Satzlängenverteilung in einen synergetisch-linguistischen Kontext (vgl. Köhler, 1986) stellt und dabei konkret systeminterne und systemexterne Einflussfaktoren in Betracht zieht. Bei einer derartigen Modellierung steht vor allem die Frage im Vordergrund, inwiefern der Autor, die Textsorte, der Stil, der Rezipient u.ä. einen Einfluss auf die Adäquatheit eines theoretischen Verteilungsmodells haben.

Satzlängenforschungen sind aus den genannten Gründen in der Gegenwart nach wie vor ein aktuelles und wichtiges interdisziplinäres Thema. Im Gegensatz zu früheren Untersuchungen werden gegenwärtige Satzlängen-Analysen jedoch in der Regel (a) korpusbasiert und (b) computergestützt durchgeführt. Gerade im Hinblick auf den zuletzt genannten Punkt ist jedoch in erster Linie eine elementare Rahmenbedingung derartiger Untersuchungen zu hinterfragen: nämlich, inwiefern operationale Kriterien einer Satzdefinition postuliert werden können, die überhaupt für eine computerbasierte Analyse zugänglich sind. Aus diesem Grunde wird im Folgenden als erstes (a) näher auf die Frage einer möglichen automatisierten Bestimmung der Satzlänge in Texten einzugehen sein (wobei wir uns hier exemplarisch auf slowenische Prosatexte beschränken), um dann in einem zweiten Schritt (b) zu prüfen, inwiefern sich signifikante Unterschiede bei der quantitativen Auswertung der Satzlänge auf der Ebene der Mittelwerte, der Schiefe und der Kurtosis ergeben, und dann in einem dritten Schritt (c) zu analysieren, inwiefern die Wahl einer bestimmten Satzdefinition als ein Einflussfaktor der theoretischen Modellierung von Satzlängen anzusehen ist.

3 Automatisierte Bestimmung der Einheit ‚Satz'

Die Frage, was unter einem ‚Satz' zu verstehen ist, stellt in der Satzlängenforschung einen durchaus prominenten Aspekt dar. Die für verschiedene theoretische und praktische Fragestellungen notwendige Definition von ‚Satz' ist allerdings keineswegs einheitlich. Dabei sollte es eigentlich als Gemeinplatz anzusehen sein, dass für jegliche formale und quantitative Arbeiten eine stringente und intersubjektiv nachvollziehbare Bestimmung der jeweiligen Untersuchungseinheit(en) – im gegebenen Fall also des ‚Satzes' und der für die Bestimmung der Satzlänge notwendigen Maßeinheiten – notwendig ist; vgl. dazu die theoretischen Implikationen der Quantifizierung und Formalisierung von Altmann und Lehfeldt (1980).

[2] Eine ausführliche Diskussion zu adäquaten Verteilungsmodellen der Häufigkeitsverteilungen der Satzlänge findet sich in Kelih und Grzybek (2004)

Während aus systemlinguistischer Sicht die Messung der Satzlänge in der Anzahl der Teilsätze pro Satz plausibel erscheint, erweist sich aus pragmatischen Gründen die Berechnung der Wortlänge in der Anzahl der Wörter pro Satz als ein mindestens ebenso praktikabler Weg, weil in diesem Fall keine zusätzlichen syntaktischen Analysen notwendig sind. In Anbetracht der Tatsache, dass offenbar beide Vorgangsweisen zum Nachweis sprachlicher Regularitäten führen, wird in der vorliegenden Arbeit davon ausgegangen, dass die Anzahl der Wörter pro Satz eine durchaus sinnvolle Maßeinheit ist. Das ‚Wort' seinerseits wird dabei auf orthographischer Ebene definiert und als eine durch eine Leerstelle abgegrenzte Einheit des Textes verstanden – zur Diskussion alternativer Wortdefinitionen und deren Auswirkung auf quantitative Untersuchungen (vgl. Antić et al., 2006).

Im vorliegenden Text soll es also weniger um eine Untersuchung der Maßeinheit gehen, in welcher die Länge eines Satzes zu berechnen ist, sondern vielmehr um die Definition von ‚Satz'. Im Grunde genommen geht es also um die primäre Frage, ob und wie sich aufgrund von klar definierten Kriterien die (linguistische) Einheit ‚Satz' in Texten und Korpora automatisiert bestimmen lässt. Aus dieser primären Fragestellung leiten sich zwei weitere ab, die mit der Auswirkung der jeweiligen Entscheidung in Zusammenhang stehen, nämlich:

1. Sind in Abhängigkeit von der Wahl einer bestimmten Satzdefinition statistisch signifikante Unterschiede in der durchschnittlichen Satzlänge, der Schiefe und der Kurtosis der Satzlängenverteilungen in den untersuchten Texten zu beobachten?

2. Hat die Wahl der Satzdefinition einen Einfluss auf postulierte theoretische Wahrscheinlichkeitsverteilungen?

Zumindest innerhalb der quantitativen bzw. statistischen Linguistik besteht relativ große Einigkeit darüber, dass Interpunktionszeichen die Funktion haben, einen schriftlichen Text in Einheiten zu gliedern. Daher können Interpunktionszeichen wie der Satzendepunkt herangezogen werden, um Satzgrenzen in Texten zu bestimmen. Eine bekannte Definition, die zahlreichen konkreten analytischen Arbeiten zugrunde liegt, ist die von Bünting und Bergenholtz (1995, p. 27); ihnen zufolge sind Sätze in Texten "solche Einheiten, die durch Satzzeichen eingegrenzt sind". In einer ersten Annäherung handelt es sich hier um eine sinnvolle Operationalisierung. Ungeachtet allfälliger Fragen der Interpunktion in Abhängigkeit von Editionsproblemen[3] ist im hier gegebenen Stelle die Frage nach dem Inventar von Interpunktionszeichen, welche als Markierung eines expliziten Satzendes

[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 sehen ist, dass die Interpunktion aufgrund von Editionseingriffen nicht immer der Autorenintention (vgl. Wake, 1957, p. 334) entspricht. So zeigte Janson (1964) in einer akribischen Studie, dass unterschiedliche Editionen in der Setzung der Interpunktionszeichen stark divergieren. Dies heißt aber auch, dass beispielsweise die durchschnittliche Satzlänge aufgrund unterschiedlicher Editionen in einem beträchtlichen Maß divergiert. Im Grunde genommen ist damit jedoch nur ein Teilproblem jeder empirischen Untersuchung angesprochen und impliziert eine eindeutige Darlegung der jeweils verwendeten Textbasis.

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 des interpunktorischen Satzes verwendet werden. Während in zahlreichen, vor allem auch früheren Arbeiten zum Deutschen wie z.B. derjenigen von Weiß (1968, p. 55), ausschließlich der Punkt, das Frage- und das Ausrufezeichen als satzabschließende Interpunktionszeichen gewertet wurden, finden sich in neueren Arbeiten wie z.B. bei Niehaus (1997, p. 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 in weiterer Folge natürlich, dass solche Definitionsunterschiede nicht ohne Auswirkung auf quantitative Berechnungen bleiben. Und um diese Frage soll es im vorliegenden Text gehen; zunächst aber gilt es, bevor wir derartige Definitionen automatisch auf die zu analysierenden slowenischen Texte übertragen, die Spezifik der Setzung von Interpunktionszeichen in den Texten der erwähnten Sprache aufzuzeigen.

3.1 Satzdefinition und automatische Bestimmung der Satzlänge im Slowenischen

Nach den Regeln der slowenischen 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 (vgl. Pravopis, 1990, p. 38ff.). Unter Berücksichtigung der Wichtigkeit der erwähnten Interpunktionszeichen bei der Textgliederung lautet somit eine erste – in der Praxis durchaus gängige und bei entsprechenden Analysen angewandte Arbeitsdefinition:

Definition 1 *Ein Satz ist eine durch Punkt, Frage- und 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 der automatischen Analyse der Texte ‚manipuliert' wird, unter Berücksichtigung der Tatsache, dass der Punkt in der Funktion als Kennzeichnung von Abkürzungen (Beispiele: *c.kr., sv.*) und in der Form von Aufzählungen (Beispiel: *To je Stritarjevo ... tam*) nicht als in satzabschließender Position vorkommend gezählt wird. Während die Problematik der Abkürzungen sich gegebenenfalls durch eine automatisierte Auflösung in Form eines Abkürzungsverzeichnisses überwinden lässt (was aufgrund der starken morphologischen Variationen im Slowenischen bei weitem nicht unproblematisch ist), bleibt die Schwierigkeit des Umgangs mit durch mehrere Punkte gekennzeichneten Aufzählungen bestehen. Noch problematischer an der Definition 1 ist jedoch, dass der Punkt, das Fragezeichen und das Ausrufezeichen nicht in allen Fällen unbedingt das Ende eines Satzes kennzeichnen müssen. Aus diesem Grund wird im folgenden gezeigt, dass auch eine alternative Zählung möglich und sinnvoll ist.

So bietet es sich in Anlehnung an die grundsätzlichen Überlegungen von Grinbaum (1996) zur Automatisierung von Satzlängenuntersuchungen beispielsweise an, den Großbuchstaben als weiteres satzabgrenzendes Zeichen in eine formal bestimmbare Satzdefinition einzubauen. Wenn auch der Großbuchstabe im Slowenischen primär 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, p. 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 (Textende, Absatzende), gelangt man zu der folgenden alternativen operationalen Satzdefinition 2.

Definition 2 *Als Satzendezeichen gelten [.], [...], [?] und [!], es sei denn, ein Kleinbuchstabe ist das erste Graphem des nächsten Wortes.*

Ohne Frage ist es möglich, mit beiden aufgezeigten Satzdefinitionen, den Satz und damit auch die Satzlänge automatisiert bestimmen zu können. Die beiden vorgestellten Satzdefinitionen stellen den Ausgangspunkt für unsere empirische Untersuchung der Satzlängenverteilung in slowenischen Texten dar. In weiterer Folge wird dann zu prüfen sein, ob die Anwendung von unterschiedlichen Satzdefinitionen Auswirkung auf statistische Kenngrößen wie Mittelwert, Schiefe und Kurtosis hat. A priori ist bei der Anwendung der Satzdefinition 2 zu erwarten, dass sich die absolute Anzahl der Sätze gegenüber Satzdefinition 1 verringert. Der Grund dafür ist darin zu sehen, dass auch Ausrufe- und Fragesätze innerhalb eines Satzgefüges als vollwertige Sätze betrachtet werden, wie das folgende Beispiel aus Text #1[4] zeigt: *"Kako se vam godi, oče – sedaj, ko ste za starega?" vprašal sem ga s smehom.* Nach Satzdefinition 1 werden hier zwei Sätze mit zehn und fünf Wörtern gezählt, während aufgrund von Satzdefinition 2 ein einziger Satz mit 15 Wörtern ausgewertet wird. Ähnlich wird beispielsweise die folgende Sequenz aus Text #3.2 je nach Satzdefinition entweder als ein Satz mit neun Wörtern oder als zwei Sätze mit fünf und vier Wörtern ausgezählt.: *"Kaj jo je zbodlo, ženščuro?" se je začudil Jernej [...]"*

3.2 Textkorpus der slowenischen literarischen Texte

Als Basis für die empirische Untersuchung dient ein Korpus slowenischer Prosatexte. Dieses setzt sich aus sechs Kurzromanen bzw. Kurzgeschichten (slowenisch: povest), im Folgenden als ‚Kurzerzählungen' bezeichnet, von vier verschiedenen slowenischen Autoren aus dem 19. Jhd. zusammen (vgl. Tabelle 1).

Diese Texte werden im Folgenden mit den entsprechenden Nummern (Text #1 bis Text #6) bezeichnet und als solche analysiert. Zum Zwecke der Kontrolle der Datenbasis

[4] Die hier genannten arabischen Zahlen bezeichnen die im Kapitel 3.2 eingeführte Nummerierung der analysierten Texte.

Text	Autor	Titel
# 1	J. Kersnik	*Mačkova očeta*
# 2	J. Kersnik	*Ponkrčev oča*
# 3	I. Cankar	*Hlapec Jernej in njegova pravica*
# 4	J. Jurčič	*Nemški Valpet*
# 5	F. Levstik	*Pokljuk*
# 6	F. Levstik	*Martin Krpan*

Tabelle 1: Das Korpus slowenischer Texte.

– zur Frage der Homogenität in quantitativen Untersuchungen vgl. Altmann (1992) bzw. Orlov (1982) – werden diese Texte jedoch nicht nur jeweils einzeln als komplexe Gesamttexte analysiert, sondern auf zwei weitere Arten und Weisen: Zum einen werden die genannten sechs Texte zu einem Gesamtkorpus zusammengefügt, so dass sich eine umfangreichere Textmischung ergibt; dieses Gesamtkorpus soll im folgenden bedingt als "Text #7" bezeichnet werden (vgl. Tabelle 2). Zum anderen ergibt sich aufgrund der Tatsache, dass die Texte #2 und #3 jeweils aus mehreren Kapiteln bestehen, die Option, diese einzelnen Kapitel jeweils als homogene Texte zu verstehen und getrennt zu analysieren; in diesem Fall haben wir es mit den Texten #8 bis #28 zu tun (vgl. Tabelle 3). Auf diese Art und Weise lässt sich die Qualität des den Analysen zugrunde gelegten Datenmaterials zuverlässig kontrollieren. In Übersicht lässt sich nunmehr die Satzlänge auf den folgenden drei Ebenen bestimmen:

1. Auf der ersten Ebene werden die genannten Kurzerzählungen (vgl. Tabelle 1) zu einem vollständigen Korpus zusammengefasst. Das Korpus, welches in Hinsicht auf die involvierten Textsorten als homogen zu bezeichnen ist, gibt die Möglichkeit zu prüfen, inwiefern eine Korpusanalyse gegebenenfalls eine andere Satzlängenverteilung aufweist als die Analyse der einzelnen Texte. Insgesamt besteht das Korpus aus 39016 Wörtern; gemäß Satzdefinition 1 werden insgesamt 2938 Sätze ausgezählt, während bei Anwendung von Satzdefinition 2 insgesamt 2758 Sätze zu verbuchen sind (also ein Unterschied von immerhin ca. 6.5%). Die weiteren Werte, vor allem auch die mittlere Satzlänge, sind im Einzelnen der Tabelle 2 zu entnehmen.

Satzdefinition	Text	Sätze	Wörter	$\bar{x}$
1	#7	2938	39016	13,28
2		2758	39016	14,15

Tabelle 2: Quantitative Angaben zum Textkorpus.

2. Auf der zweiten Ebene werden die sechs Kurzerzählungen als komplexe Texte aufgefasst; anzumerken ist, dass dabei eine textinterne, von den Autoren selbst vorgenommene Kapitelgliederung nicht beachtet wird. Insgesamt handelt es sich also um sechs unterschiedliche Kurzerzählungen von vier verschiedenen Autoren, wobei Text #3 mit insgesamt 1383 Sätzen den größten Umfang aufweist. Unter dieser Voraussetzung zeichnen sich die Texte durch die in Tabelle 3 zusammengefassten Charakteristika aus.

		Satzdefinition 1		**Satzdefinition 2**	
Text	**Wörter**	**Sätze**	$\bar{x}$	**Sätze**	$\bar{x}$
# 1	1597	118	13,53	109	14,65
# 2	2178	191	11,40	165	13,20
# 3	18407	1493	12,33	1383	13,31
# 4	7971	585	13,63	561	14,21
# 5	3181	170	18,71	169	18,82
# 6	5682	381	14,91	371	15,32

Tabelle 3: Quantitative Angaben zu den komplexen Texten.

3. Auf der dritten Ebene werden die Texte unter Berücksichtigung der von den Autoren selbst vorgenommen Kapitelgliederung jeweils individuell untersucht. Im Detail weist der Text von Janko Kersnik *Ponkrčev oča* drei Kapitel auf; Ivan Cankars *Hlapec Jernej in njegova pravica* besteht aus 18 Einzelkapiteln; insgesamt stehen also 21 Einzelkapitel für die Analysen zur Verfügung. Tabelle 4 resümiert die wesentlichen Charakteristika der Einzeltexte.

Durch die differenzierte Analyse auf drei unterschiedlichen Ebenen ergeben sich insgesamt 28 Datensätze, für welche die Satzlänge unter Anwendung beider Satzdefinitionen bestimmt werden kann. Auf der Basis dieser Texte lässt sich nunmehr – neben der Frage des Einflussfaktors der Satzdefinition – auch die Frage der Datenhomogenität kontrollieren.

3.3 Statistische Vergleiche der durchschnittlichen Satzlänge

Als erstes wird in den sechs komplexen Texten (vgl. Tabelle 1) nach Satzdefinitionen 1 und 2 die in der Anzahl der Worte gemessene Satzlänge automatisiert bestimmt. Es zeigt sich in diesem ersten Schritt, dass aufgrund der Satzdefinition 2 in allen Texten eine geringere absolute Anzahl an Sätzen ($N_1 < N_2$) ausgezählt wird. Dieser Befund deutet darauf hin, dass aufgrund der Satzdefinition 2 der Text in größere (d.h. längere) Einheiten eingeteilt wird; dementsprechend ändert sich auch die in den Texten berechnete durchschnittliche Satzlänge. Beispielsweise ergibt sich für den komplexen Text #2 aufgrund der Satzdefinition 1 eine mittlere Satzlänge von $\bar{x} = 11{,}40$ Wörtern

		Satzdefinition 1		Satzdefinition 2	
Text	**Wörter**	**Sätze**	**$\bar{x}$**	**Sätze**	**$\bar{x}$**
# 8	895	76	11,78	68	13,16
# 9	523	44	11,89	38	13,76
# 10	760	71	10,70	59	12,88
# 11	602	47	12,81	43	14,00
# 12	977	92	10,62	82	11,91
# 13	1038	90	11,53	85	12,21
# 14	796	61	13,05	57	13,96
# 15	809	81	9,99	80	10,11
# 16	890	81	10,99	75	11,87
# 17	973	79	12,32	71	13,70
# 18	1473	120	12,28	107	13,77
# 19	939	65	14,45	60	15,65
# 20	1134	120	9,45	113	10,04
# 21	937	80	11,71	75	12,49
# 22	1203	80	15,04	80	15,04
# 23	1583	126	12,56	119	13,30
# 24	956	61	15,67	53	18,04
# 25	1388	107	12,97	98	14,16
# 26	1203	84	14,32	77	15,62
# 27	1203	98	12,28	87	13,83
# 28	303	21	14,43	21	14,43

Tabelle 4: Quantitative Angaben zu den Einzeltexten.

pro Satz, während auf der Grundlage von Satzdefinition 2 die mittlere Satzlänge $\bar{x}$ = 13,20 beträgt. Man sieht somit, dass die beiden Satzdefinitionen sich offensichtlich unmittelbar und massiv auf die durchschnittliche Satzlänge auswirken.

Es stellt sich nun die Frage, inwiefern diese Beobachtung an allen Texten nachzuweisen ist, oder ob die soeben angesprochenen Unterschiede als ein Einzelfall zu betrachten sind. Entsprechend werden für alle slowenischen Datensätze die durchschnittlichen Satzlängen aufgrund der beiden angeführten Satzdefinitionen berechnet, und die sich ergebenden Satzlängenunterschiede auf statistische Signifikanz geprüft; die einzelnen Werte finden sich in der Tabelle 5.

Wie zu sehen ist, unterscheiden sich lediglich zwei der 28 Stichproben nicht im Hinblick auf die durchschnittliche Satzlänge (Text #22 und Text #28). Bei allen anderen Datensätzen wirkt sich die Satzdefinition auf den Mittelwert aus, wobei sich natürlich die Frage nach der Signifikanz des Unterschiedes stellt. Für einen statistischen Vergleich der Mittelwerte unter beiden Bedingungen – die wir als unabhängige Stichproben ansehen wollen – ist es üblich, den sog. Zweistichproben t-Test durchzuführen,

der auch in der Linguistik breite Verwendung gefunden hat. Dieser t-Test ist u.a. in der Satzlängenforschung angewendet worden, um zu testen, ob sich zwei Mittelwerte aus zwei unabhängigen Stichproben auf einem festgelegten Niveau unterscheiden (vgl. Grzybek, 2000, p. 446). Geprüft wird also die Nullhypothese, dass sich zwei Mittelwerte auf einem festgelegten Signifikanzniveau ($\alpha = 0{,}05$ und fakultativ $\alpha = 0{,}01$) nicht unterscheiden. Die Berechnungsverfahren des t-Werts unterscheiden sich geringfügig in Abhängigkeit davon, ob die Varianzen beider Stichproben homogen sind oder nicht, was mit Hilfe der sog. Levene-Statistik berechnet werden kann; in unserem Fall zeigt die Levene-Statistik, dass alle Stichproben Homogenität der Varianzen aufweisen. Als Ergebnis dieses Mittelwertvergleichs stellt sich heraus, dass sich lediglich zwei der 28 Stichproben – nämlich Text #3, der mit Abstand der längste der Texte ist, sowie Text #7, das gesamte Korpus) – auf dem 1%-Niveau signifikant unterscheiden. Ein weiterer Text (Text #2) kommt hinzu, wenn man das Signifikanzniveau bei 5% ansetzt; hierbei handelt es sich interessanterweise um einen solchen Text, der einen hohen Anteil an Frage- und Ausrufesätzen aufweist.

Bevor man jedoch zu vorschnellen Interpretationen gelangt, gilt es folgendes zu berücksichtigen: Voraussetzung für die Durchführung des t-Tests ist jedoch, dass die Werte beider Stichproben normalverteilt sind; dies ist in unseren beiden Stichproben – wie entsprechende Tests zeigen – allerdings nicht der Fall. Da somit die Anwendung des t-Tests aufgrund der Verletzung der vorausgesetzten Normalverteilung nicht zulässig ist, muss der sog. U-Test nach Mann/Whitney zum Einsatz kommen. Ähnlich wie der t-Test dient auch er dem Vergleich von zwei Stichproben hinsichtlich ihrer zentralen Tendenz, allerdings können hier die Werte beliebig verteilt sein oder Ordinalniveau aufweisen. Tabelle 5 enthält in den beiden letzten Spalten die entsprechenden z-Werte sowie die ihnen entsprechenden Wahrscheinlichkeiten.

Wie der Tabelle 5 zu entnehmen ist, führt die zulässige Anwendung des U-Tests in unserem Fall im Wesentlichen zu ein und denselben Ergebnissen wie der t-Test: Auf dem 1%-Niveau gibt es signifikante Abweichungen nur beim gesamten Korpus (Text #7), sowie bei den Texten #2 und #3; hinzu kommt lediglich Text #10, wenn man die Signifikanzschwelle auf 5% senkt.

Damit lässt sich, dieses erste Ergebnis zusammenfassend, sagen, dass die beiden vorgestellten Satzdefinitionen in der Tat zu signifikanten Unterschieden führen können. Bemerkenswert ist dabei, dass dies in erster Linie beim Gesamtkorpus (Text #7) und bei dem längsten Text #3 der Fall ist; insofern scheint es plausibel anzunehmen, dass eine bestimmte Anzahl von Beobachtungen (d.h. eine gewisse Textlänge) notwendig ist, damit überhaupt signifikante Unterschiede zum Tragen kommen können. Andererseits scheint es aber durchaus auch textspezifische bzw. textsortenspezifische Einflüsse zu geben, die in einem signifikanten Mittelwertunterschied resultieren können.

Ohne Zweifel ist die Frage nach der zentralen Tendenz einer Stichprobe bzw. nach Unterschieden in der zentralen Tendenz eine der meist gestellten Fragen im Rahmen von Satzlängenforschungen. Der Grund für die Beliebtheit dieser Fragestellung dürfte darin zu sehen sein, dass es hier um die Spezifik individueller Texte geht. In einem breiteren Kontext jedoch scheint eine in eine andere Richtung zielende Frage von mindestens

	n		$\bar{x}$		s		t-Test			Levene-Test		U-Test	
Text	n_1	n_2	$\bar{x}_1$	$\bar{x}_2$	s_1	s_2	t	FG	p		p	z	p
#7	2938	2758	13,28	14,15	9,82	9,79	-3,33	5694	**0,001**	0,168	0,682	-4,48	< **0,001**
#1	118	109	13,53	14,65	9,95	10,48	-0,82	225	0,411	0,010	0,921	-0,90	0,3660
#2	191	165	11,40	13,20	8,44	8,30	-2,02	354	**0,044**	0,137	0,712	-2,65	**0,0080**
#3	1493	1383	12,33	13,31	9,00	8,94	-2,93	2874	**0,003**	0,288	0,592	-3,81	< **0,001**
#4	585	561	13,63	14,21	9,58	9,65	-1,03	1144	0,310	0,042	0,837	-1,24	0,2160
#5	170	169	18,71	18,82	11,89	11,87	-0,09	337	0,932	0,000	0,982	-0,10	0,9200
#6	381	371	14,91	15,32	11,62	11,56	-0,48	750	0,635	0,002	0,960	-0,75	0,4530
#8	76	68	11,78	13,16	8,27	7,98	-1,02	142	0,309	0,143	0,706	-1,28	0,2000
#9	44	38	11,89	13,76	7,76	8,24	-1,06	80	0,291	0,077	0,782	-1,31	0,1920
#10	71	59	10,70	12,88	9,06	8,81	-1,38	128	0,170	0,147	0,702	-2,06	**0,0390**
#11	47	43	12,81	14,00	8,66	8,41	-0,66	88	0,510	0,045	0,832	-0,92	0,3560
#12	92	82	10,62	11,91	8,00	8,05	-1,06	172	0,289	0,030	0,862	-1,34	0,1800
#13	90	85	11,53	12,21	8,34	8,35	-0,54	173	0,592	0,008	0,927	-0,73	0,4660
#14	61	57	13,05	13,96	15,31	15,49	-0,32	116	0,747	0,000	0,977	-0,73	0,4680
#15	81	80	9,99	10,11	7,46	7,47	-0,11	159	0,916	0,001	0,979	-0,13	0,8950
#16	81	75	10,99	11,87	7,83	7,75	-0,70	154	0,482	0,046	0,831	-1,01	0,3140
#17	79	71	12,32	13,70	8,27	7,98	-1,04	148	0,289	0,095	0,758	-1,37	0,1710
#18	120	107	12,28	13,77	8,01	8,11	-1,39	225	0,165	0,010	0,919	-1,54	0,1240
#19	65	60	14,45	15,65	7,66	6,92	-0,92	123	0,360	0,741	0,391	-0,90	0,3670
#20	120	113	9,45	10,04	7,17	7,24	-0,62	231	0,536	0,002	0,966	-0,71	0,4790
#21	80	75	11,71	12,49	7,20	6,94	-0,69	153	0,493	0,213	0,645	-0,75	0,4550
#22	80	80	15,04	15,04	11,28	11,28	0,00	158	1,000	0,000	1,000	0,00	1,0000
#23	126	119	12,56	13,30	8,78	8,59	-0,67	243	0,506	0,062	0,804	-0,84	0,3990
#24	61	53	15,67	18,04	11,60	11,46	-1,09	112	0,277	0,015	0,904	-1,55	0,1200
#25	107	98	12,97	14,16	8,33	8,07	-1,04	203	0,301	0,085	0,771	-1,20	0,2290
#26	84	77	14,32	15,62	10,40	10,12	-0,80	159	0,423	0,083	0,773	-1,16	0,2440
#27	98	87	12,28	13,83	7,37	6,98	-1,47	183	0,145	0,223	0,637	-1,53	0,1260
#28	21	21	14,43	14,43	9,35	9,35	0,00	40	1,000	0,000	1,000	0,00	1,0000

Tabelle 5: Vergleich der durchschnittlichen Satzlängen.

ebenso großer Bedeutung; diese geht ebenfalls von bestimmten Charakteristika der Häufigkeitsverteilung aus, fragt jedoch in erster Linie nach einem allfällig gemeinsamen Profil der Verteilungen.

3.4 Statistischer Vergleich von Schiefe und Kurtosis

Konkret bietet sich zur Untersuchung der zuletzt genannten Fragestellung die Analyse von Schiefe (γ_1) und Kurtosis (γ_2) der Häufigkeitsverteilung an. Mit diesen beiden Maßen wird angegeben, in welchem Maße eine Verteilung im Vergleich zur Normalverteilung links- oder rechtsverschoben bzw. höher oder niedriger als diese liegt: Im Falle einer linkssteilen Verteilung spricht man von einer positiven Schiefe (d.h. $\gamma_1 > 0$), im Fall einer steilgipfligen Verteilung von einem positiven Exzeß ($\gamma_2 > 0$). Im gegebenen Fall interessiert zwar nicht in erster Linie der Vergleich zur Normalverteilung, wohl aber die Frage, ob es signifikante Unterschiede in Schiefe und/oder Kurtosis in Abhängigkeit von der Satzdefinition gibt.

Tabelle 6 repräsentiert die Werte für Schiefe und Kurtosis für beide Satzdefinitionen.

Das Maß der Schiefe (γ_1) ist hier nach der üblichen Formel 1 berechnet:

$$\hat{\gamma}_1 = \frac{m_3}{s^3}, \tag{1}$$

	Schiefe		Vergleich Schiefe		Kurtosis		Vergleich Kurtosis	
Text	γ_1 (1)	γ_1 (2)	z	p	γ_2 (1)	γ_2 (2)	z	p
# 7	1,9490	1,9760	-0,2811	0,7787	8,0370	8,2700	-0,2465	0,8053
# 1	1,4082	1,6671	-1,6975	0,0896	3,3184	4,3692	-1,5431	0,1228
# 2	1,2154	1,1316	1,1147	0,2650	1,2154	1,0473	0,7077	0,4791
# 3	2,0300	2,0545	-0,1486	0,8818	9,8870	10,4402	-0,3480	0,7278
# 4	1,6374	1,6401	-0,0419	0,9666	3,6903	3,5506	0,3954	0,6925
# 5	1,4115	1,4196	-0,0897	0,9286	3,0785	3,0960	-0,0469	0,9626
# 6	2,2759	2,3054	-0,1176	0,9064	10,6192	10,8907	-0,1305	0,8962
# 8	0,9217	0,8320	0,7574	0,4488	0,5303	0,5648	-0,0993	0,9207
# 9	1,1929	1,1195	0,6075	0,5435	0,8089	0,3235	1,2163	0,2239
# 10	1,4996	1,4117	0,8015	0,4228	1,9920	1,7811	0,4494	0,6531
# 11	1,5894	1,6429	-0,4517	0,6515	2,7709	2,8772	-0,1649	0,8690
# 12	1,1366	1,1049	0,3240	0,6400	0,8356	1,2083	-1,1318	0,2577
# 13	1,6953	1,6086	0,4955	0,6203	4,5089	4,3323	0,2159	0,8291
# 14	3,4854	3,4589	0,1139	0,9093	14,8248	14,3772	0,1739	0,8619
# 15	1,2904	1,2632	0,2246	0,8223	1,8174	1,7635	0,1255	0,9001
# 16	1,3642	1,3839	-0,2066	0,8363	1,4508	1,4708	-0,0489	0,9610
# 17	1,1457	1,1000	0,4475	0,6545	1,1984	1,2190	-0,0595	0,9525
# 18	0,8086	0,7935	0,1810	0,8564	0,3222	0,8042	-1,9601	0,0500
# 19	0,3429	0,5021	-1,8844	0,0595	-0,5813	-0,4817	-0,5844	0,5590
# 20	1,3580	1,2329	0,7955	0,4263	2,9518	2,6195	0,6302	0,5285
# 21	0,5450	0,5093	0,4039	0,6863	0,0256	0,1825	-0,8115	0,4171
# 22	2,4774	2,4774	0,0000	1,0000	10,5769	10,5769	0,0000	1,0000
# 23	1,2186	1,2452	-0,1933	0,8468	2,2737	2,4774	-0,3469	0,7287
# 24	2,2453	2,2475	-0,0115	0,9908	7,7227	7,7275	-0,0052	0,9958
# 25	0,6540	0,5516	1,5894	0,1120	-0,3293	-0,3797	0,3800	0,7039
# 26	2,0087	2,1433	-0,6821	0,4952	6,2466	6,9015	-0,6421	0,5208
# 27	0,5747	0,4985	1,1295	0,2587	-0,1940	-0,1980	0,2662	0,9788
# 28	0,8701	0,8701	0,0000	1,0000	0,5659	0,5659	0,0000	1,0000

Tabelle 6: Kennwerte zur Schiefe und Kurtosis.

wobei $s^2 = \frac{N}{N-1} \sum_{i=1}^{k} (i - \bar{x})^2 p_i$.

Wie der Tabelle 6 zu entnehmen – und wie nicht anders zu erwarten – ist das Maß der Schiefe in allen Texten unter beiden Bedingungen jeweils positiv. Um nun das Ausmaß der Schiefe für beide Satzdefinitionen miteinander zu vergleichen, ist es notwendig, gemäß der Formel 2 die Maße für die Schiefe und die Varianz der Schiefe unter beiden Bedingungen – in der Formel als (a) und (b) gekennzeichnet – zueinander in Beziehung zu setzen.

$$z_1 = \frac{\hat{\gamma}_{1(a)} - \hat{\gamma}_{1(b)}}{\sqrt{Var\left(\hat{\gamma}_{1(a)}\right) + Var\left(\hat{\gamma}_{1(b)}\right)}} \qquad (2)$$

Während die Varianz der Schiefe üblicherweise unter Annahme der Normalverteilung der Werte nach der Formel 3 berechnet wird,

$$Var\left(\hat{\gamma}_1\right) = \frac{6N \cdot (N-1)}{(N-2) \cdot (N+1) \cdot (N+3)}, \tag{3}$$

ist bei fehlender (bzw. nicht anzunehmender) Normalverteilung die Berechnung nach Lewis und Orav (1989) etwas komplizierter gemäß Formel 4 vorzunehmen:

$$Var\left(\hat{\gamma}\right) = \frac{(N-1)\, L\left(\hat{\gamma}\right)}{4\,(N-2)^2\,(s)^{10}} \tag{4}$$

wobei $L\left(\hat{\gamma}\right) = 4s^4m_6 - 12s^2m_3m_5 - 24s^6m_4 + 9\left(m_3\right)^2 m_4 + 35s^4\left(m_3\right)^2 + 36s^{10}$.

Als Ergebnis der entsprechenden Vergleiche stellt sich heraus, dass sich in allen Fällen die Schiefe nicht signifikant in Abhängigkeit von der jeweiligen Satzdefinition unterscheidet. Dasselbe gilt für die Kurtosis (γ_2), deren Werte ebenfalls der Tabelle 6 zu entnehmen sind. Die Kurtosis wird nach der üblichen Formel 5 berechnet:

$$\hat{\gamma}_2 = \frac{m_4}{s^4} - 3 \tag{5}$$

Zur Prüfung, ob sich das Maß der Kurtosis für beide Satzdefinitionen unterscheidet, ist es abermals notwendig, die Maße für die Kurtosis und die Varianz der Kurtosis unter beiden Bedingungen – in der Formel wiederum als (a) und (b) gekennzeichnet – zueinander in Beziehung zu setzen, und zwar gemäß der Formel 6.

$$z_2 = \frac{\hat{\gamma}_{2(a)} - \hat{\gamma}_{2(b)}}{\sqrt{Var\left(\hat{\gamma}_{2(a)}\right) + Var\left(\hat{\gamma}_{2(b)}\right)}} \tag{6}$$

Die Varianz der Kurtosis berechnet sich hierbei nach der Formel 7:

$$Var\left(\hat{\gamma}_2\right) = \frac{(N-1)^2\left(N^2-2N+3\right)^2 L\left(\hat{\gamma}_2\right)}{(N-2)^2\,(N-3)^2\,(N)^3\,(s)^{12}} \tag{7}$$

wobei $L\left(\hat{\gamma}_2\right) = s^4m_8 - 4s^2m_4m_6 - 8s^4m_3m_5 + 4{m_4}^3 - s^4{m_4}^2 + 16s^2{m_3}^2m_4 + 16s^6{m_3}^2$.

Damit stellt sich insgesamt heraus, dass im Gegensatz zum Befund signifikanter Mittelwertunterschiede (s.o.), in allen Texten die Satzdefinition weder bei der Schiefe noch bei der Kurtosis einen signifikanten Unterschied bewirkt. Dabei gibt es unter beiden Bedingungen einen positiven linearen Zusammenhang zwischen Schiefe und Kurtosis; dies zeigt der aufgrund der teilweise fehlenden Normalverteilung der Variablen zu berechnende Spearman'sche Rangkorrelationskoeffizient ($\rho = 0.98$, für Satzdefinition 1 bzw. $\rho = 0.97$ für Satzdefinition 2, jeweils $p < 0.001$).

Abbildung 1 veranschaulicht das Ergebnis für das Gesamtkorpus (#Text 7), welches auf graphischer Ebene deutlich macht, dass sich das Profil der Häufigkeitsverteilungen in der Tat nicht wesentlich in Abhängigkeit von der Satzdefinition ändert.

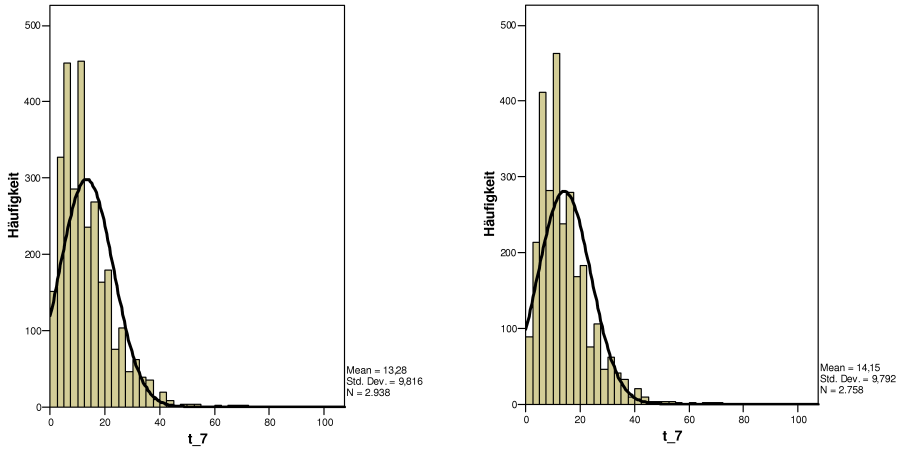

Abbildung 1: Satzlängenhäufigkeitsverteilung für das gesamte Korpus (Text #7) mit eingezeichneter Normalverteilungskurve.

Diese Beobachtung leitet allerdings zu der weiterführenden Frage über, inwiefern sich an die Texte ein einheitliches Modell einer diskreten Wahrscheinlichkeitsverteilung anpassen lässt, und inwiefern sich hier entweder im Hinblick auf das Modell insgesamt oder aber in Hinsicht auf die Parameterwerte des entsprechenden Modells Unterschiede in Abhängigkeit von der Satzdefinition ergeben.

4 Theoretische Modellierungen von Satzlängen

Bei der Verfolgung dieser Fragestellung können wir uns auf Ergebnisse einer anderen Studie beziehen, deren Design hier nicht im Detail dargestellt werden muss (vgl. Kelih und Grzybek, 2004). Es ging in dieser Studie um einen anderen möglichen Faktor aus dem Umfeld der Rahmenbedingungen, der möglicherweise die theoretische Modellierung der Satzlängenhäufigkeit beeinflusst: nämlich das zum Zwecke der Datenglättung üblicherweise angewendete Verfahren der Intervallbildungen, das aufgrund der mit Satzlängenhäufigkeiten in der Regel verbundenen hohen Streuung notwendig ist. Ohne auf die Ergebnisse dieser Untersuchung hier im Detail einzugehen, können wir festhalten, dass sich bei den oben genannten Texten (vgl. Tabellen 2-4) die negative Binomialverteilung als durchgehend robustes Modell am besten zur theoretischen Modellierung der Satzlängenhäufigkeiten erweisen hat. Vor diesem Hintergrund erweist es sich nunmehr im Hinblick auf die beiden oben diskutierten Satzdefinitionen als sinnvoll, die allfällige Auswirkung der Satzdefinition auf dieses Modell bzw. dessen konkrete Parameter zu überprüfen.

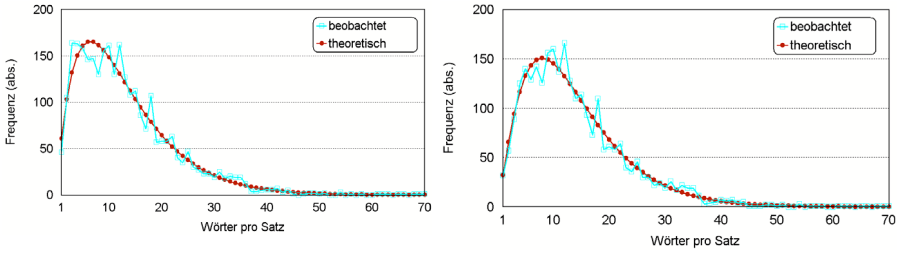

Abbildung 2: Anpassung der negativen Binomialverteilung an die Satzlängenhäufigkeitsverteilung für das gesamte Korpus (Text #7) gemäß Satzdefinition 1 und 2.

4.1 Empirische Überprüfung der Verteilung von Satzlängen in slowenischen Texten

Die negative Binomialverteilung ist gegeben als

$$P_x = \left(\begin{array}{c} k+x-1 \\ x \end{array} \right) p^k q^x \qquad x = 0, 1, 2, ... \tag{8}$$

Da Sätze mit 0 Wörtern jedoch prinzipiell ausgeschlossen sind, wird sie korrekterweise in ihrer 0-gestutzten Form angewendet, wie sie in (9) gegeben ist:

$$P_x = \left(\begin{array}{c} k+x-1 \\ x \end{array} \right) \frac{p^k q^x}{1-p^k} \qquad x = 1, 2, ... \tag{9}$$

Die Anpassung dieser Verteilung an die Daten lässt sich mit Spezialsoftware wie dem Altmann-Fitter (2000) computergestützt erreichen (der auch die 0-Stutzung automatisch vollzieht). Dabei werden die Werte für die Parameter nach bestimmten Anfangsschätzungen in iterativen Prozeduren so optimiert, dass der χ^2-Wert als Test für die Güte der Anpassung minimiert wird. Veranschaulichen wir das Vorgehen an Text #7, dem Gesamtkorpus.

Abbildung 2 stellt die beobachteten Häufigkeiten für beide Satzdefinitionen dar, die in Form eines zusammenfassenden Histogramms schon in Abbildung 1 dargestellt wurden (s.o.). Hinzu kommen nun die theoretischen Häufigkeiten, wie sie sich nach Einsetzen der Werte für die Parameter k und p ergeben: Für Satzdefinition 1 betragen die Werte $k_1 = 1.97$ und $p_1 = 0.14$, für Satzdefinition 2 betragen $k_2 = 2.41$ und $p_2 = 0.16$. Das Einsetzen dieser Werte in die obige Formel resultiert in einem Wert von $\chi^2 = 53.78$ bzw. $\chi^2 = 44.47$ für die beiden Satzdefinitionen. Dies entspricht für Satzdefinition 1 einer Wahrscheinlichkeit von $P = 0.007$ bei 25 Freiheitsgraden, für Satzdefinition 2 von $P = 0.0067$ (bei 24 Freiheitsgraden). Da üblicherweise Wahrscheinlichkeiten im Falle von $P > 0.05$ als sehr gutes bzw. $P > 0.01$ als gutes Anpassungsergebnis zu werten sind, wäre hier von einer schlechten Anpassung zu sprechen. Allerdings steigt der χ^2-Wert linear mit der Stichprobengröße, weshalb er für große Stichproben Abweichungen schneller als signifikant erscheinen lässt; aus diesem Grund wird in der quantitativen Linguistik im Falle

von großen Stichproben statt der Wahrscheinlichkeit P der Wert des Diskrepanzkoeffizienten C herangezogen, der sich als $C = \chi^2 \ / \ N$ berechnet (vgl. Grotjahn und Altmann, 1993). Dies bietet sich im Falle von Text #7 bei Stichprobenumfängen von $N_1 = 2938$ bzw. $N_2 = 2758$ an und resultiert in als gut zu bezeichnenden Anpassungswerten von $C_1 = 0.0183$ bzw. $C_2 = 0.0161$. Fasst man die Satzlängenhäufigkeit in 5er-Intervallen zusammen – wie dies aufgrund der hohen Streuung in der Satzlängenforschung absolut üblich ist – resultiert dies in einer Verringerung der Werte im Falle von Satzdefinition 1 auf $C = 0.0121$ und bei Satzdefinition 2 auf $C = 0.0048$, was sich als Indiz für die ausgezeichnete Eignung der negativen Binomialverteilung interpretieren lässt.

Tabelle 7 stellt für alle Texte die Ergebnisse der Anpassung der negativen Binomialverteilung für beide Satzdefinitionen dar; enthalten ist neben dem jeweiligen χ^2-Wert mit den dazugehörigen Freiheitsgraden (FG) die diesem Wert entsprechende Wahrscheinlichkeit P, den sich aus der Division von χ^2 und N ergebenden C-Wert sowie die jeweiligen Parameterwerte.

	Satzdefinition 1					***Satzdefinition 2***				
Text	χ^2	FG	P	C	N	χ^2	FG	P	C	N
# 7	53,78	25	0,0007	**0,0183**	2938	44,47	24	0,0067	**0,0161**	2758
# 1	30,55	35	**0,6829**	0,2589	118	38,04	36	**0,3767**	0,349	109
# 2	36,16	32	**0,2803**	0,1893	191	32,98	33	**0,4680**	0,1999	165
# 3	24,23	5	0,0002	**0,0162**	1493	11,85	4	0,0185	**0,0086**	1383
# 4	58,89	44	**0,0660**	0,1007	585	7,52	5	**0,1845**	0,0134	561
# 5	39,03	41	**0,5583**	0,2296	170	42,15	42	**0,4646**	0,2494	169
# 6	59,87	46	**0,0823**	0,1571	381	15,20	8	**0,0555**	0,041	371
# 8	28,11	26	**0,3529**	0,3699	76	21,74	27	**0,7505**	0,3197	68
# 9	22,06	18	**0,2294**	0,5013	44	21,94	18	**0,2347**	0,5774	38
# 10	18,72	22	**0,6623**	0,2637	71	0,58	3	**0,9002**	0,0099	59
# 11	24,84	21	**0,2541**	0,5286	47	23,19	18	**0,1834**	0,5393	43
# 12	23,39	25	**0,5550**	0,2542	92	13,84	25	**0,9644**	0,1688	82
# 13	23,03	25	**0,5757**	0,2559	90	25,14	26	**0,5113**	0,2957	85
# 14	33,75	24	**0,0892**	0,5533	61	10,41	6	**0,1085**	0,1826	57
# 15	30,30	21	**0,0861**	0,3741	81	27,74	22	**0,1846**	0,3467	80
# 16	24,35	23	**0,3846**	0,3006	81	19,41	23	**0,6772**	0,2588	75
# 17	17,98	25	**0,8434**	0,2275	79	20,78	25	**0,7047**	0,2927	71
# 18	31,22	29	**0,3552**	0,2602	120	38,10	29	**0,1201**	0,3561	107
# 19	3,58	7	**0,8265**	0,0551	65	4,08	6	**0,6664**	0,0679	60
# 20	18,97	13	**0,1241**	0,1581	120	41,17	28	**0,0518**	0,3643	113
# 21	35,37	24	**0,0631**	0,4421	80	23,55	17	**0,1321**	0,3141	75
# 22	19,14	28	**0,8937**	0,2393	80	19,14	28	**0,8937**	0,2393	80
# 23	21,33	29	**0,8469**	0,1693	126	9,03	28	**0,8972**	0,1599	119
# 24	30,44	28	**0,3427**	0,4989	61	17,93	26	**0,8782**	0,3384	53
# 25	25,55	27	**0,5435**	0,2388	107	24,82	28	**0,6377**	0,2533	98
# 26	22,88	28	**0,7389**	0,2724	84	22,17	28	**0,7735**	0,2879	77
# 27	28,16	26	**0,3506**	0,2874	98	21,68	24	**0,5982**	0,2492	87
# 28	10,73	13	**0,6333**	0,511	21	10,73	13	**0,6333**	0,511	21

Tabelle 7: Anpassungsergebnisse der negativen Binomialverteilung.

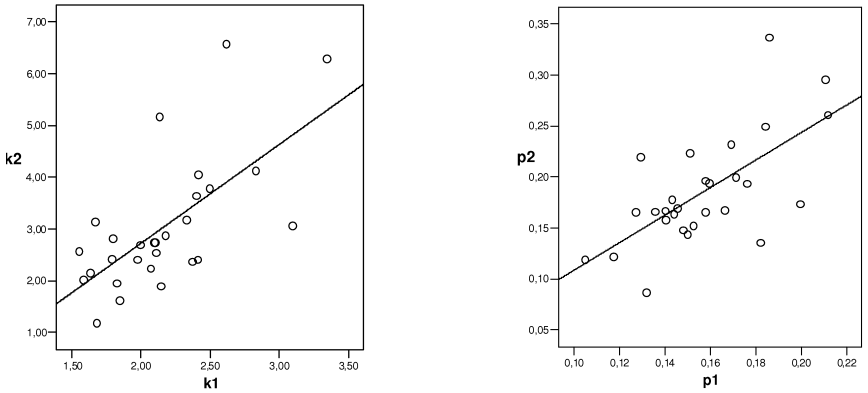

Abbildung 3: Parameter k und p der negativen Binomialverteilung für Satzdefinition 1 und 2.

Wie zu sehen ist, erweist sich die negative Binomialverteilung unter der Bedingung beider Satzdefinitionen als geeignetes Modell: Sowohl für Satzdefinition 1 als auch für Satzdefinition 2 lassen sich die Texte durch ein und dasselbe theoretische Verteilungsmodell, nämlich die negative Binomialverteilung, beschreiben. Dabei sind keinerlei Unterschiede in Bezug auf die analysierte Textebene erkennbar, das Modell erweist sich sowohl auf der Ebene der Einzeltexte, als auch der komplexen Texte, als auch des Korpus unter beiden Bedingungen als hervorragend geeignet.

Ungeachtet dessen stellt sich die einen Schritt weiter gehende Frage, ob sich die Wahl der Satzdefinition auf die theoretischen Parameterwerte der Verteilung auswirkt.

5 Parameter der negativen Binomialverteilung

Im Hinblick auf die beiden Parameter der negativen Binomialverteilung (k, p) gibt es eine Reihe möglicher Fragen, was einen eventuellen Einfluss der Satzdefinitionen betrifft. Eine erste Frage zielt darauf, ob die beiden Parameter jeweils eine systematische Verschiebung in Abhängigkeit von der Satzdefinition erfahren.

Wie Abbildung 3 zeigt, gibt es einen positiven linearen Zusammenhang sowohl für k als auch für p in Abhängigkeit von der Satzdefinition: k und p tendieren dazu, unter der Bedingung von Satzdefinition 2 größer zu sein als bei der Anwendung von Satzdefinition 1; die Tendenz ist in beiden Fällen gleichermaßen hoch signifikant, wie der aufgrund der teilweise fehlenden Normalverteilung der Variablen zu berechnende Spearman'sche Rangkorrelationskoeffizient zeigt ($\rho = .63$ bzw. $\rho = .62$ für die Parameter k und p, jeweils $p < 0.001$). Eine etwaige Abhängigkeit der Parameter von der (in der Anzahl der Sätze gemessenen) Textlänge ist dabei nicht erkennbar.

	Satzdefinition 1			Satzdefinition 2		
Text	k_1	p_1	N	k_2	p_2	N
# 7	1,9746	0,1403	2938	2,4109	0,1582	2758
# 1	1,5810	0,1048	118	2,0227	0,1190	109
# 2	1,5507	0,1272	191	2,5731	0,1656	165
# 3	1,7865	0,1356	1493	2,4157	0,1660	1383
# 4	2,0949	0,1431	585	2,7412	0,1781	561
# 5	3,0963	0,1499	170	3,0636	0,1440	169
# 6	1,8216	0,1174	381	1,9531	0,1218	371
# 8	1,7964	0,1401	76	2,8140	0,1672	68
# 9	2,8281	0,2117	44	4,1330	0,2612	38
# 10	1,6654	0,1510	71	3,1466	0,2236	59
# 11	2,6144	0,1858	47	6,5782	0,3371	43
# 12	1,6329	0,1440	92	2,1535	0,1637	82
# 13	2,0681	0,1663	90	2,2424	0,1674	85
# 14	1,6774	0,1319	61	1,1899	0,0867	57
# 15	2,1435	0,1995	81	1,8966	0,1737	80
# 16	2,1066	0,1761	81	2,5422	0,1937	75
# 17	2,3279	0,1712	79	3,1800	0,1996	71
# 18	1,9943	0,1456	120	2,6939	0,1697	107
# 19	3,3410	0,2106	65	6,2973	0,2959	60
# 20	1,8440	0,1820	120	1,6238	0,1359	113
# 21	2,4949	0,1841	80	3,7872	0,2497	75
# 22	2,4091	0,1524	80	2,4091	0,1524	80
# 23	2,1748	0,1596	126	2,8750	0,1944	119
# 24	2,1318	0,1292	61	5,1725	0,2197	53
# 25	2,0984	0,1579	107	2,7400	0,1656	98
# 26	2,4008	0,1578	84	3,6458	0,1964	77
# 27	2,4113	0,1690	98	4,0521	0,2318	87
# 28	2,3706	0,1481	21	2,3706	0,1481	21

Tabelle 8: Die Parameter k und p der negativen Binomialverteilung.

Allerdings ist in Bezug auf die Parameter k und p ein weiterer interessanter Zusammenhang zu beobachten, auf den in der Geschichte der Satzlängenforschung bislang noch nicht aufmerksam gemacht worden ist, und den es in Zukunft detailliert und systematisch zu verfolgen gilt. Dieser Zusammenhang betrifft eine auffällige und signifikante Abhängigkeit des Parameters p vom Parameter k: üblicherweise allerdings wird bei der Interpretation derartiger Zusammenhänge nicht der Parameter p herangezogen, sondern der sich als $1 - p$ berechnende Parameter q der negativen Binomialverteilung.

Abbildung 4 veranschaulicht den linearen Zusammenhang zwischen den Parametern q

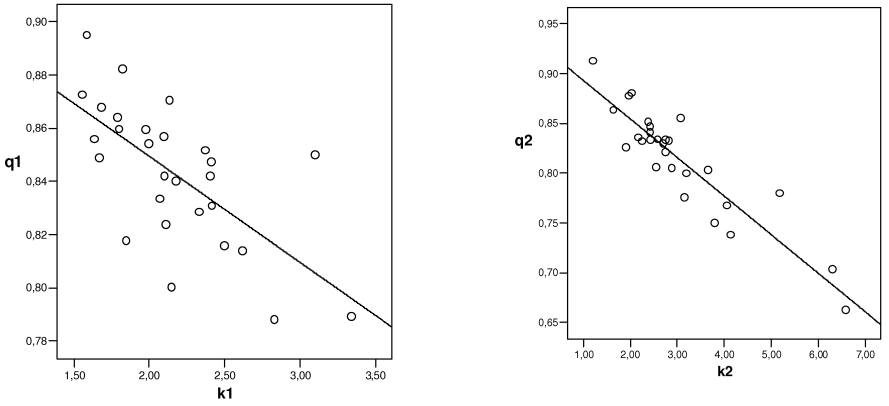

Abbildung 4: Abhängigkeit des Parameters q vom Parameters k der negativen Binomialverteilung für Satzdefinition 1 und 2.

und k für beide Satzdefinitionen. Wie deutlich zu sehen ist, ist der Zusammenhang für Satzdefinition 2 wesentlich stärker; das bestätigt der aufgrund der teilweise fehlenden Normalverteilung zu berechnende Spearman'sche Rangkorrelationskoeffizient, der im Falle von Satzdefinition 1 einen Wert von $\rho = -.70$ aufweist ($p < 0.001$), im Falle von Satzdefinition 2 einen Wert von $\rho = -.91$ ($p < 0.001$).

Abgesehen davon, dass dieser Zusammenhang als ein starkes Argument für die Güte der Satzdefinition 2 angesehen werden kann, ergibt sich für die Theorie der Satzlängenforschung an dieser Stelle die Aufgabe, den Zusammenhang zwischen den beiden Parametern q und k der negativen Binomialverteilung einer qualitativen Interpretation auf der Grundlage umfangreicheren Datenmaterials zu unterziehen.

6 Resümee

Abschließend gilt es, die erhaltenen Resultate kurz zusammenzufassen:

1. Diskutiert wurde die bislang in der Satzlängenforschung vernachlässigte Fragestellung der Anwendung und Auswirkung unterschiedlicher Satzdefinitionen; diese Diskussion ist als zentral für jegliche weiterführende Untersuchungen zu sehen. Im vorliegenden Fall konnte gezeigt werden, dass die Anwendung von zwei (qualitativ prinzipiell als gleichwertig anzusehenden) Satzdefinitionen zu statistisch signifikanten Unterschieden des Mittelwertes führen kann. Als statistische Verfahren wurden dabei der (im gegebenen Fall die zu seiner Durchführung notwendigen Voraussetzungen nicht erfüllende) t-Test und der Mann-Whitney'sche U-Test angewandt. Es konnte gezeigt werden, dass nicht mehr als ca. 15% der Texte einen signifikanten

Unterschied in den Mittelwerten aufweisen. Erklärbar ist dies einerseits durch die Stabilität der beiden Satzdefinitionen, andererseits durch die gewählte homogene Textbasis (ausschließlich slowenische Prosatexte). Beide angewandten Tests zeigen insbesondere auf der Ebene des Gesamtkorpus und bei dem längsten Text signifikante Unterschiede, was den Schluss nahelegt, dass signifikante Unterschiede erst bei einer größeren Anzahl von Satzlängen mit unterschiedlicher Länge zum Tragen kommen.

2. Aufgrund der eingeschränkten Aussagekraft von Mittelwerten (die aber dennoch eine in der Satzlängenforschung sehr beliebte Kenngröße sind) wurden auch die Profile der sich aufgrund der Satzdefinitionen ergebenden Satzlängenverteilungen einem statistischen Signifikanztest in Form eines Vergleich von Kurtosis und Schiefe unterworfen. Es zeigt sich hierbei, dass für keinen der Texte ein signifikanter Unterschied nachgewiesen werden kann.

Insofern würde man bereits a priori feststellen können, dass auch die Anwendung von unterschiedlichen Satzdefinitionen auf der Ebene der theoretischen Modellierung keinen bzw. nur geringen Einfluss haben kann. Dieser Befund konnte durch entsprechende empirische Überprüfungen bestätigt werden, in denen die negative Binomialverteilung als geeignetes theoretisches Modell der Satzlängenverteilung in allen zugrunde gelegten slowenischen Prosatexten nachgewiesen werden konnte. Darüber hinaus lassen sich auf der Ebene der theoretischen Modellierung folgende Resultate anführen:

1. Die beiden Satzdefinitionen lassen einen Zusammenhang erkennen, der sich an der negativen Korrelation der jeweiligen aus den theoretischen Verteilungen resultierenden Parameterwerte (k_1 und k_1 bzw. p_1 und p_2) nachweisen lässt. Insofern lässt sich postulieren, dass die Anwendung der beiden angewandten Satzdefinitionen zu einer *systematischen Verschiebung der Parameter* führt.

2. Als wichtig und richtungsweisend anzusehen ist auch der Befund eines statistischen Zusammenhangs zwischen den beiden Parametern k und p der negativen Binomialverteilung. Die beiden Parameter sind bei beiden Satzdefinitionen hoch korreliert, wobei der deutlich ausgeprägtere Zusammenhang bei Satzdefinition 2 ($\rho = -.91$) als ein Argument dafür zu interpretieren ist, dass die Satzdefinition 2 als die qualitativ adäquatere Satzdefinition anzusehen ist.

Insgesamt ist dieser Beitrag als eine zentrale Detailuntersuchung der Satzlängenforschung zu sehen, wobei eine Ausweitung auf weitere Textsorten und Sprachen wünschenswert wäre. Darüber hinaus kann mit der vorgestellten Satzdefinition systematisch die Frage untersucht werden, inwiefern die Satzlänge und daraus berechnete Kenngrößen ein adäquates Mittel für eine Klassifizierung von Texten, Textsorten bzw. Funktionalstilen herangezogen werden kann (vgl. Kelih et al., 2006).

Literatur

Altmann, G. (1988a). Verteilungen der Satzlängen. In K. P. Schulz (Hrsg.), *Glottometrika 9*, S. 147–161. Bochum: Brockmeyer.

Altmann, G. (1988b). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Altmann, G. (1992). Das Problem der Datenhomogenität. In B. Rieger (Hrsg.), *Glottometrika 9*, S. 287–298. Bochum: Brockmeyer.

Altmann, G. und W. Lehfeldt (1980). *Einführung in die Quantitative Phonologie*. Bochum: Brockmeyer.

Antić, G., E. Kelih und P. Grzybek (2006). *Contributions to the Science of Language. Word Length Studies and Related Issues*, Kapitel *Zero-syllable Words in Determining Word Length*, S. 117–156. Dordrecht, NL: Springer.

Bolton, H. C. und A. Roberts (1995). On the comparison of literary and scientific styles: The letters and articles of Max Born, FRS. *Notes and Records of the Royal Society of London 49*(2), 295–302.

Bünting, K. D. und H. Bergenholtz (1995). *Einführung in die Syntax*. Stuttgart: Beltz.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology 32*, 221–233.

Grinbaum, O. N. (1996). *Prikladnoe jazykoznanie*, Kapitel *Komp'juternye aspekty stilemetrii*, S. 451–463. Sankt-Peterburg: Izd. S.-Peterburgskogo Univ.

Grotjahn, R. und G. Altmann (1993). *Contributions to Quantitative Linguistics*, Kapitel *Modelling the Distribution of Word Length: Some Methodological Problems*, S. 141–153. Dordrecht: Kluwer Academic.

Grzybek, P. (2000). *Slovo vo vremeni i prostranstve. K 60-letiju professora V. M. Mokienko*, Kapitel *Zum Status der Untersuchung von Satzlängen in der Sprichwortforschung – Methodologische Vor-Bemerkungen*, S. 430–457. Moskva: Folio-Press.

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities 28*, 87–106.

Janson, T. (1964). The problems of measuring sentence-length in classical texts. *Studia Linguistica 18*, 26–36.

Karlgren, J. und D. Cutting (1994). Recognizing text genres with simple metrics using discriminant analysis. In M. Nagao (Hrsg.), *Proceedings of COLING 94*, S. 1071–1075.

Kelih, E. und P. Grzybek (2004). Häufigkeiten von Satzlängen: Zum Faktor der Intervallgröße als Einflussvariable (am Beispiel slowenischer Texte). In *Glottometrics 8*, S. 23–41. Brockmeyer.

Kelih, E., P. Grzybek, E. Stadlober und G. Antić (2006). *Text Classification: The Impact of Sentence Length*. Heidelberg: Springer.

Kjetsaa, G. (1984). *The Authorship of the Quiet Don*. Oslo: Solum Forl.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Lewis, P. A. und E. J. Orav (1989). *Simulation Methodology for Statisticians, Operations Analysts, and Engineers*. Pacific Grove, CA: Wadsworth & Brooks.

Mistrík, J. (1973). Eine exakte Typologie von Texten. In *Arbeiten und Texte zur Slavistik*, Band 3. München: Verlag Otto Sagner.

Niehaus, B. (1997). Untersuchung zur Satzlängenhäufigkeit. In K.-H. Best (Hrsg.), *Glottometrika 16*, S. 213–276. Trier: WVT.

Orlov, J. K. (1982). Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie „Sprache-Rede" in der statistischen Linguistik). In J. K. Orlov, M. G. Boroda und I. Š. Nadarejšvili (Hrsg.), *Sprache, Text, Kunst. Quantitative Analysen*, S. 1–55. Bochum: Brockmeyer.

Pieper, U. (1979). *Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse*. Tübingen: Narr.

Pravopis (1990). *Slovenski pravopis. I: Pravila*. Ljubljana: Državna založba Slovenije.

Sherman, L. A. (1888). Some observations upon the sentence-length in English prose. In *Studies, 1*, S. 119–130. University of Nebraska.

Sichel, H. S. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society (A) 137*, 25–34.

Sigurd, B., M. Eeg-Olofsson und J. van de Weijer (2004). Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica 58*(1), 37–52.

Smith, M. W. A. (1983). Recent experience and new developments of methods for the determination of authorship. *Bulletin of the Association for Literary and Linguistic Computing 11*(3), 73–82.

Teigeler, P. (1968). *Verständlichkeit und Wirksamkeit von Sprache und Text*. Stuttgart: Verlag Nadolski.

Tuldava, J. (1993). Measuring text difficulty. In G. Altmann (Hrsg.), *Glottometrika 14*, S. 69–81. Bochum: Brockmeyer.

Wake, W. C. (1957). Sentence-length distributions of Greek authors. *Journal of the Royal Statistical Society 120*, 331–346.

Weiß, H. (1968). *Statistische Untersuchungen über Satzlänge und Satzgliederung als autorenspezifische Stilmerkmale. (Beitrag zur mathematischen Analyse der Formalstruktur von Texten)*. Doktorarbeit, Aachen, T. H., Math.-naturwiss. F.

Williams, C. B. (1940). A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika 31*, 356–361.

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose. *Biometrika 30*, 363–390.

Software

Altmann-Fitter (2000): *Iterative Fitting of Probability Distributions*. Lüdenscheid: RAM Verlag.