Zur Prosa Karel Čapeks – Einige quantitative Bemerkungen

Peter Grzybek, Ernst Stadlober (Graz)

Im Jahr 1934 – im selben Jahr also, in dem Jan Mukařovský seinen klassischen Text «L'art comme fait sémiologique» für den VIII. Internationalen Philosophenkongreß in Prag konzipierte – gab er eine für den Schulgebrauch gedachte Sammlung mit Prosatexten von Karel Čapek heraus *(Výbor z prózy Karla Čapka*). Diese erschien 1946 in zweiter Auflage; sie enthielt Erzählungen, Novellen, Essays, journalistische Texte, sowie Ausschnitte aus verschiedenen Romanen.

Die Herausgabe von Čapek-Texten durch Mukařovský ist ein interessantes Thema für sich, das eine Reihe von Fragen aufwirft: Findet hier etwa Mukařovskýs Suche nach der noetischen Dimension des Kunstwerks eine künstlerische Entsprechung in Čapeks (von ihm selbst so bezeichneten) „noetischen Erzählungen"? Oder hat diese Konstellation eher soziokulturelle Bedeutung: Steht sie vielleicht im Zusammenhang mit der Diskussion, die der Prager Linguistische Kreis in den 30er Jahren insbesondere mit der Redaktion von *Naše řeč* zur Sprachkultur und Norm der Literatursprache führte? Denn in dieser Diskussion figurierte u.a. auch Čapek 1935 mit seinem kurzen Text «Wenn ich ein Linguist wäre» (*Kdybych byl lingvistou*), und zwar prominent in der ersten Nummer von *Slovo a slovesnost,* unmittelbar nach der programmatischen Einleitung der Herausgeber. Čapek akzentuierte hier die sprachliche Individualität des Schriftstellers; Einheitlichkeit, Regulation und allgemeine Ordnung seien eine Sache der Abstraktion. Nicht uninteressant ist in diesem Zusammenhang der Umstand, daß Mukařovský dem genannten Sammelband eine ausführliche Darstellung der „Entwicklung von Karel Čapeks Prosa" (*Vývoj Čapkovy prózy*) voranstellte, die 1948 dann auch in seinen *Kapitoly z české poetiky* wiederabgedruckt wurde. Darin gliederte Mukařovský das Schaffen Čapeks in folgende Phasen, die er mit „Veränderungen in der künstlerischen Komposition" (56) begründete.

1. Eine erste Phase sah Mukařovský in den 1916 bzw. 1918 erschienenen Bänden *Krakonošova zahrada* («Rübezahls Garten»; im folgenden: *KZ*) und *Zářivé hlubiny* («Leuchtende Tiefen»; im folgenden: *ZH*). Mukařovský betonte als gemeinsame Merkmale beider Bände „die Art des Stils" (56); einen Grund dafür sah er in der Zusammenarbeit der Brüder Čapek an den Texten dieser beiden Bände sowie im Bezug zur literarischen Tradition: So sei *KZ* gekennzeichnet durch das parodistische „Mißverhältnis zwischen Sprachausdruck und Inhalt" (58); doch schon die Erzählungen in *ZH* seien durch den „Übergang von der lyrischen zur erzählenden Prosa" (59) gekennzeichnet und stellten insofern eine „Schule des epischen Erzählens" (61) dar. Hier löse eine „radikale Vereinfachung des Stils" (59) sowie „Sparsamkeit und Einfachheit des Wortausdrucks" (60) ein kompliziertes, aus Para- und Hypotaxe bestehendes System syntaktischer Abhängigkeiten (57) ab.

2. In eine zweite Phase ordnete Mukařovský die 1917 bzw. 1921 erschienenen Erzählsammlungen *Boží muka* («Die Gottesmarter»; im folgenden: *BM*) und *Trapné povídky* («Peinliche Geschichten»; im folgende *TP*) ein, ebenso wie die Romane *Továrna na absolutno* («Die Fabrik für das Absolute») von 1922 und *Krakatit* (1924). Beginnend mit *BM* handelt es sich nur noch um selbständige Werke von Karel Čapek. Typisch sei hier die Tendenz, auf die *Art der Darbietung*, auf den sprachlichen Ausdruck aufmerksam zu machen, technisch realisiert durch eine Zweiplanigkeit der Darstellung (73): im Vordergrund stehe eine anti-realistisch motivierte Betonung des Unterschieds von Wirklichkeit und Nachricht durch die doppelte Spiegelung von Ereignissen.

3. Die Tendenz der zweiten Phase verstärke sich dann in der dritten Phase, in die Mukařovský die *Povídky z jedné kapsy* («Die Erzählungen aus der einen Tasche»; im folgenden: *1K*) (1929) sowie die Romane *Hordubal* (1933) und *Povětroň* («Der Meteor») (1934) einord-

net. Dominant sei hier die „Technik der mehrfachen Spiegelung ein und derselben vorausgesetzten Wirklichkeit“ (82).[1]

Obwohl sich Mukařovský zeitgleich in seinen theoretischen Schriften nachdrücklich für die Betrachtung der formalen Elemente eines Kunstwerks als Träger einer potentiellen semantischen Energie aussprach, argumentiert er hier vornehmlich auf inhaltlicher Ebene. Vor diesem Hintergrund sollen im folgenden formale Aspekte der genannten Texte einer detaillierteren Analyse unterzogen werden; Ziel der Untersuchung ist es somit, die in der Geschichte der Text- und Literaturwissenschaft immer wieder postulierte dialektische Wechselbeziehung von Form und Inhalt einer operationalisierbaren Beschreibung zu unterziehen.

Bei dem Versuch, dieses im Rahmen des vorliegenden Aufsatzes bestenfalls ansatzweise realisierbare Ziel zu erreichen, wird man ohne Frage nicht um die Anwendung mathematisch-statistischer Verfahren herumkommen. Dabei gelten noch heute unverändert zwei Bemerkungen, mit denen Helmut Kreuzer schon 1965 den von ihm gemeinsam mit Rul Gunzenhäuser herausgegebenen Band *Mathematik und Dichtung* einleitete: nämlich erstens, daß jeder Versuch zur Ausbreitung von Mathematisierungs- und Formalisierungstendenzen in Bereiche, die ihnen bisher entzogen waren, auf ein kulturkritisches Mißtrauen stößt; und zweitens, daß es eine reizvolle Aufgabe wäre, das in verschiedenen Epochen wechselnde Interesse an solchen Mathematisierungen zu untersuchen. Solche wechselnden und gewissermaßen widersprüchlichen Tendenzen können durchaus zeitlich parallel existieren; in dem uns vertrauten derzeitigen Wissenschaftsgefüge sind im Bereich der Kulturwissenschaften ohne Frage eher ‘anti-mathematische’ Modelle dominant. Andererseits aber kommen literaturtheoretische Arbeiten, die unter dem Einfluß der Chaos-Theorie entstanden sind – man vergleiche nur Lotmans wichtigen Aufsatz „Über die Rolle des Zufälligen in der Literatur“ – nicht ohne den Faktor des Zufalls aus, was zwangsläufig zu wahrscheinlichkeitstheoretischen Modellen und zu Fragen der Berechenbarkeit des Zufalls führt, wofür Bände wie z.B. der unlängst erschienene zum Thema *Künste des Zufalls* (Gendolla/Kamphusmann 1999) ein beredtes Beispiel sind.

Eine tiefe Verwurzelung haben quantitative Untersuchungen allerdings auch und gerade in der Tradition des tschechoslowakischen Strukturalismus. Der Blick zurück reicht mitunter bis ins 17. Jahrhundert, nämlich bis auf Aussagen von Jan Amos Komenský (1592–1670) zur Ökonomisierung des Fremdsprachenunterrichts auf der Basis von Wortvorkommenshäufigkeiten. Und von mehr als anekdotischer Bedeutung ist der Beitrag von Seydler, der 1885/86 die Berücksichtigung der Wahrscheinlichkeitstheorie in die Diskussion um die Autorschaft des 1817 gefundenen *Rukopis Královédvorský a Zelenohorský* einbrachte. Prominenten Platz nahmen quantitative Fragen dann in den ‘klassischen’ Arbeiten von Mathesius, Trnka, Vachek oder Krámský ein – Arbeiten, an die man in den 60er Jahren anknüpfte. Die für die damaligen Arbeiten charakteristische enge Verbindung zwischen quantifizierender und allgemeiner Linguistik setzt sich heute fort – die Querverbindungen zur Literaturwissenschaft hingegen sind bei weitem nicht so eng, wie dies Mitte

[1] Aus der an und für sich chronologisch argumentierten Typologie gliedert Mukařovský eine Reihe von Texten wie z.B. die *Povídky z druhé kapsy* («Erzählungen aus der anderen Tasche») aus, die er explizit aus dem chronologischen Zusammenhang herausnimmt.

der 60er Jahre der Fall war, als zahlreiche quantitative Arbeiten zur Verstheorie und zur poetischen Sprache erschienen. Diese im Grenzbereich von Literatur- und Sprachwissenschaft angesiedelten Untersuchungen etwa von Lubomír Doležel oder Jiří Levý bezogen sich auf frühere Arbeiten zur Dichtersprache, verstanden als eine Funktionssprache im Gegensatz zur Mitteilungssprache. Aus diesem Gegensatz wurden schließlich die antagonistischen Tendenzen der Automatisierung und Aktualisierung abgeleitet, die freilich ihren Ursprung in der Verfremdungsästhetik des Russischen Formalismus haben. Die Entwicklung der Dichtersprache ist demnach durch das Streben nach Deformation charakterisiert – ein Streben gegen die Mitteilungssprache einerseits, gegen erstarrte poetische Schemata andererseits. So steht die Dichtersprache in einem ständigen ('dynamischen') Spannungsfeld zweier gegenläufiger Kräfte – der Aktualisierung und der Normierung.

In der informationstheoretischen Euphorie der 60er Jahre nahm man an, daß die quasi-statistischen Aussagen zur Dichtersprache leicht transformiert und in eine statistische Theorie eingegliedert werden könnten – so etwa Doležel (1965, 278). Doležel (1967) schrieb der Theorie der Dichtersprache *in toto* einen prä-statistischen Charakter zu: die relevanten Grundaussagen seien zwar nicht der Art ihrer Formulierung, wohl aber ihrem Inhalt nach statistischer Natur.

Dennoch führten diese Arbeiten in eine Sackgasse – und das nicht nur, weil sich der Strukturalismus als Auslaufmodell erweisen sollte. Ein anderer Grund war, daß man damals üblicherweise von der Existenz bestimmter Normen und Standards (sei es der Sprache insgesamt oder bestimmter sprachlicher Register) ausging. Von der kumulativen Untersuchung möglichst vieler Texte und Textkorpora zielte man auf Standards der Graphem- und Worthäufigkeit, der Wort- und Satzlänge und anderes mehr ab; diese sollten als konventionelle Folie für allfällige Deviationen dienen. So gab Doležel zum Beispiel für das Tschechische allgemein eine durchschnittliche Satzlänge von $\bar{x}$ = 16,03 Wörter pro Satz an; ergänzend führte er spezifische Werte für verschiedene stilistische Bereiche an, die in Tab. 1 wiedergegeben sind.

Tab. 1: Mittlere Satzlänge(n) in tschechischen Funktionalstilen (nach Doležel 1967)

Drama	$\bar{x}$ = 4,56
Gedichte	$\bar{x}$ = 14,26
Zeitungen	$\bar{x}$ = 14,61
Naturwissenschaftl. Texte	$\bar{x}$ = 18,87
Gesetzestexte	$\bar{x}$ = 27,40

Deutlich erkennbar unterscheidet sich die durchschnittliche Satzlänge bei den einzelnen Texttypen, ebenso offensichtlich liegen die Werte für Gedichte und Zeitungstexte – im Unterschied zu den anderen Werten – sehr nah beieinander. Ganz offensichtlich ist also Satzlänge ein mögliches texttypendifferenzierendes Kriterium, reicht aber andererseits als alleiniges Kriterium nicht für differenzierende Aussagen aus.

Richtigerweise postulierte Doležel deshalb auch die Berücksichtigung weiterer Kriterien. Gerade in dieser Hinsicht ist deshalb eine Untersuchung von Ludvíková (1972) zur Wortlänge im Tschechischen von unmittelbarer Relevanz: In dieser Studie verglich die Autorin bei Textausschnitten dreier verschiedener Texttypen im Umfang von jeweils 2.000 Wörtern die mittlere Wortlänge miteinander, nämlich Ausschnitte aus (a) einem wissenschaftlichen, mündlichen gehaltenen Vortrag, (b) einem Dialog aus einem Radiointerview, (c) literarischer Prosa von F. Hrubín. In ihrer Untersuchung wurde vorausgesetzt, daß ein Wort aus mindestens einer Silbe besteht, so daß die 0-silbigen Wörter nicht als eigenständige Wortklasse angesehen wurden. Es stellte sich heraus, daß sich in Stichproben von jeweils 2000 Wörtern aus diesen drei Texttypen die mittleren Wortlängen deutlich voneinander unterschieden: im Durchschnitt beim Dialog $\bar{x}_D$ = 1,90 Silben pro Wort, bei der literarischen Prosa $\bar{x}_P$ = 2,05 Silben pro Wort, und bei der Vorlesung $\bar{x}_V$ = 2,28 Silben pro Wort. Zwar führte Ludvíková nur die Mittelwerte ohne dazugehöriges Streuungsmaß

an; da sich jedoch im Anhang ihres Aufsatzes die kumulierten Rohdaten in tabellarischer Form finden, lassen sich nachträglich in einer Re-Analyse die Standardabweichung s und Standardfehler des Mittelwerts $s_{\bar{x}}$ berechnen. Damit ergeben sich in der Zusammenschau die in Tab. 2 angeführten Werte:

Tab. 2: Mittlere Wortlänge in verschiedenen tschechischen Textausschnitten (nach LUDVÍKOVÁ 1972)

	D*ialog*	P*rosa*	V*orlesung*
$\bar{x}$	1,90	2,05	2,28
s	0,99	0,94	1,17
$s_{\bar{x}}$	0,022	0,021	0,026

Mit den Standardfehlern des Mittelwerts lassen sich approximativ die Konfidenzintervalle berechnen: wie zu sehen ist, überschneiden sich im gegebenen Fall die 95%-Konfidenzbereiche nicht, so daß wir es mit drei heterogenen Stichproben zu tun haben. Dies bestätigt sich auch in Mittelwertvergleichen zwischen den einzelnen Stichproben: Unter der Annahme, daß die jeweiligen Stichproben aus Normalverteilungen mit unterschiedlichen Varianzen stammen, lassen sich diese in Form von t-Tests für unabhängige Stichproben durchführen; bei Mehrfachanwendung des Tests (paarweiser Vergleich mehrerer Stichproben) ist allerdings eine Korrektur des ursprünglichen Signifikanzniveaus von α = 0,05 vorzunehmen (sog. Bonferroni-Korrektur). Es stellt sich dabei heraus, daß die Unterschiede zwischen den Gruppen unter Berücksichtigung der jeweiligen Freiheitsgrade jeweils hoch signifikant ($p < 0{,}001$) sind (D–P: $t_{FG=3987} = 4{,}91$; D–V: $t_{FG=3891} = 11{,}09$; P–V: $t_{FG=3820} = 6{,}85$). Da die t-Verteilung bei Stichprobengrößen von $N \geq 120$ faktisch mit der Normalverteilung übereinstimmt, kann im gegebenen Fall der p-Wert auch über die Normalverteilung berechnet werden.

LUDVÍKOVÁ (1972) hat über die Berechnung von Mittelwerten hinausgehend versucht, ein theoretisches Verteilungsmodell an ihre Daten anzupassen. Im Zuge der damaligen Zeit unternahm sie diesen Versuch mit einer 1-verschobenen Poisson-Verteilung, wie sie auch Fucks in seinen Arbeiten als Spezialfall der von ihm beschriebenen gewichteten Poissonverteilung verwendete – jedoch ohne Erfolg. Während Fucks mit seinem Verteilungsmodell noch die Annahme verknüpft hatte, daß es übereinzelsprachlich gültig sei und für alle auf dem Silbenprinzip aufbauenden Sprachen gleichermaßen zutreffe, ist diese Annahme mittlerweile als obsolet anzusehen: Im Anschluss an eine grundlegende Erörterung methodologischer Probleme der Wortlängenmodellierung von GROTJAHN/ALTMANN (1993) wurde ein vollkommen neuer Ansatz zur Erforschung von Wortlängenhäufigkeiten entwickelt (WIMMER et al. 1994; WIMMER/ALTMANN 1996; WIMMER et al. 1999) Hierbei handelt es sich um ein flexibles System von Verteilungen: Die Grundidee besteht darin, dass die jeweils benachbarten Wahrscheinlichkeitsklassen gemäß einer einfachen Proportionalitätsbeziehung miteinander verbunden sind: $P_X \sim P_{X-1}$, d.h. die Anzahl der zweisilbigen Wörter in einem Text, steht in spezifischer Relation zur Anzahl der einsilbigen Wörter dieses Textes, die Anzahl der dreisilbigen in Relation zur Anzahl der zweisilbigen, usw. Das Verhältnis zwischen den Längenklassen erweist sich dabei nicht als konstant, sondern läßt sich als Funktion verstehen: $P_X = g(x)P_{X-1}$. In Abhängigkeit davon, welche konkrete Form $g(x)$ annimmt, kommt man folglich zu unterschiedlichen Verteilungsmodellen. Insofern ist es von Interesse, daß UHLÍŘOVÁ (1995, 1996) bei ihrer Arbeit mit tschechischen Texten mittlerweile mit Erfolg 10 Kindergeschichten von M. Macourek sowie 24 Kurzgeschichten von B. Hrabal an die Binomialverteilung bzw. an die (modifizierte) erweiterte positive Binomialverteilung angepaßt hat. Abgesehen von der andersartigen Herangehensweise – herangezogen wird nunmehr nicht nur der Mittelwert als ein Maß der zentralen Tendenz einer Häufigkeitsverteilung, sondern die Verteilung insgesamt – ist im Vergleich zu den früheren quantitativen Untersuchungen gleichzeitig auch ein ganz anderer entscheidender Schritt vollzogen: die Konzentration auf quantitative

Aspekte einzelner, vollständiger Texte. Im Fokus der Analyse steht somit jeweils ein in sich geschlossener Text (bzw., z.B. bei langen Romanen, eine jeweils in sich geschlossene Texteinheit wie Kapitel, o.ä.). Untersuchungen der Vergangenheit hingegen stützten sich häufig entweder auf willkürlich begrenzte Stichproben vollständiger Texte, d.h. auf Textauszüge (wie z.B. bei Ludvíková), oder aber auf ganze Textkorpora (vgl. Doležel). Beide Vorgehensweisen sind jedoch aus heutiger Sicht grundsätzlich zu korrigieren. Geht man nämlich davon aus, daß ein Text einem selbstregulatorischen Prozeß unterliegt, dann liegt auf der Hand, daß weder Textauszüge noch Text-Mischungen das Ergebnis dieses Prozesses adäquat widerspiegeln können. Aus dieser Einsicht heraus stellt ein jeder Text auch für die Statistik ein Individuum dar – ein Postulat, bei dem auch die traditionelle Literaturwissenschaft eigentlich hätte aufhorchen müssen…

Im Zuge dieser Einsichten konzentriert sich ein Großteil der aktuellen Forschung auf die Untersuchung von Verteilungsmodellen, wobei es u.a. um sprachtypologische, autorspezifische, genrebedingte o.ä. Charakteristika, die (a) in der Spezifik der Verteilungsmodelle und (b) in der Variation von deren Parametern gesucht werden. Im Gegensatz zu diesen 'anspruchsvolleren' Vorgehensweisen soll in den folgenden Analysen bewußt ein wesentlich simplerer Schritt nachgeholt werden, nämlich eine Überprüfung der Aussagekraft einfacher Mittelwertberechnungen bei der quantitativen Analyse einzelner Texte: Auf der Basis eines ausgewählten Textkorpus soll für jeden einzelnen Text die durchschnittliche Wortlänge berechnet werden. Als Textkorpus verwenden wir dabei – um die anfangs dargestellte Fragestellung hiermit wieder aufzugreifen – Prosatexte von Karel Čapek.

Das Textkorpus ist so zusammengesetzt, daß in Anlehnung an die oben dargestellte Periodisierung der Čapekschen Prosa durch Mukařovský Texte aus unterschiedlichen Schaffensphasen Berücksichtigung finden. Hierbei gilt es freilich zu beachten, daß die von Mukařovský genannten Sammlungen erstens Texte enthalten, die zu einem großen Teil schon früher gesondert publiziert wurden. Im einzelnen enthält das Textkorpus – eine detaillierte Aufstellung mit den Daten der Erstveröffentlichung findet sich im Anhang – 33 literarische Prosatexte, die sich fünf Kategorien zuordnen lassen, wie Tab. 3 zeigt. Dabei wurden unter Ausschluß der erwähnten Romantexte (*Továrna na absolutno; Krakatit, Hordubal*, *Povětroň)* ausschließlich Erzählungen ausgewählt, um die Außagen nicht durch eine allfällige genretypologische Heterogenität zu beeinflussen.

Tab. 3: Textkorpus der literarischen Erzählungen

	Titel des Sammelbandes	Erscheinungsjahr und -ort	Anzahl der analys. Texte	Erstveröffentl. der analysierten Texte
ZH	*Zářivé hlubiny a jiné prózy*	Praha 1916	5	1910–1913
BM	*Boží muka*	Praha 1917	5	1913–1917
KZ	*Krakonošova zahrada*	Praha 1918	13	1909–1910
TP	*Trapné povídky*	Praha 1921	5	1919–1921
1K	*Povídky z jedné kapsy*	Praha 1929	5	1928

Dieses Korpus literarischer Texte wird ergänzt durch 25 journalistische Texte aus den Jahren 1917–38; Detailangaben zu diesen Texten finden sich ebenfalls im Anhang. Somit haben wir es nach Hinzufügen der Kategorie

„Journalistische Texte (JT)“ mit einer durchaus ansehnlichen Stichprobengröße von insgesamt 58 Texten zu tun: Sie umfaßt immerhin ca. 75.000 Wörter bzw. fast 150.000 Silben.

Im Anhang finden sich für jeden der Texte die Werte für die mittlere Wortlänge mit der dazugehörigen Standardabweichung; diese Werte sind unter der Voraussetzung berechnet worden, daß 0-silbige Wörter eine eigenständige Wortlängenklasse bilden und als solche in die Mittelwertberechnung eingeflossen sind. Um allfällige Einflüsse dieser wortklassenspezifischen Behandlung zu kontrollieren, wurde für jeden Text auch die mittlere Wortlänge unter der Voraussetzung berechnet, daß die 0-silbigen keine eigene Wortklasse bilden; dies führt dazu, daß der Gesamtmittelwert sich erhöht und von $\bar{x}_0$ = 2,02 (s = 0,12) nach $\bar{x}_\varnothing$ = 2,08 (s = 0,13) verschiebt. Ein Mittelwertvergleich mit dem t-Test zeigt, daß dieser Unterschied signifikant ist ($t_{FG=113}$ = 2,50; p < 0,05), allerdings weist die Berechnung des Pearsonschen Korrelationskoeffizienten zwischen den jeweiligen Mittelwerten der einzelnen Texte bei einem Wert von r = 0.99 eine hochgradige Korrelation zwischen den Werten auf (p < 0,001), d.h. daß das Niveau sich insgesamt zwar unterscheidet, daß dies aber bei den einzelnen Werten über die gesamte Stichprobe hinweg in mehr oder weniger identischer Weise der Fall ist. Aus diesem Grunde wird in den folgenden Überlegungen weiterhin mit den Werten gerechnet, die auf der Berechnung der Wortlänge unter Berücksichtigung der 0-silbigen Wörter als eigener Wortklasse beruhen.

Auf der einen Seite steht also immerhin die stattliche Anzahl von 58 Prosatexten zur Verfügung, um zu quantifizierenden Aussagen zu gelangen; eine der möglichen Aussagen könnte z.B. die Feststellung beinhalten, daß der Unterschied der mittleren Wortlänge in Prosatexten zwischen der Untersuchung von Ludvíková (s.o.) und den soeben dargestellten Ergebnissen auf der Basis der Čapekschen Texte marginal ist. Allerdings gilt es auf der anderen Seite zu berücksichtigen, daß die von uns untersuchten Texte in sich eine nicht unwesentliche Heterogenität aufweisen, und zwar in verschiedener Hinsicht:

a. Die Texte gehen nur zum Teil auf die alleinige Autorschaft von Karel Čapek (K) zurück: Von den insgesamt 58 Texten stammen 42 Texte (72,40%) von ihm allein, 16 Texte (27,60%) sind von ihm gemeinsam mit seinem Bruder Josef signiert (JK).

b. Die Texte entstammen unterschiedlichen Funktionalstilen: 33 Texte (56,90%) sind der literarischen Prosa (LP), 25 Texte (43,10%) der journalistischen Prosa (JT) zuzuordnen.

c. Die Texte entstammen einer Zeitspanne von insgesamt ca. 30 Jahren: Der früheste wurde 1909, der späteste 1938 veröffentlicht; um einen ersten Eindruck von der Verteilung der Entstehungszeit der Texte zu gewinnen, wurde die gesamte Stichprobe am Median von 1920 geteilt; im Ergebnis dieser Teilung stammen 28 (‘frühere’) Texte (48,30%) aus den Jahren 1909 bis 1919, 30 (‘spätere’) Texte (51,70%) aus den Jahren 1920 bis 1938.

Wie der Tab. 4 leicht zu entnehmen ist, verteilt sich die Textanzahl in den einzelnen Untergruppen deutlich ungleichgewichtig:

- es gibt keine journalistischen Texte, die von Karel und Josef Čapek gemeinsam stammen würden;
- es gibt keine ‘späten’ Texte, die von Karel und Josef Čapek gemeinsam stammen würden;

• bei den 'früheren' Texten gibt es ein deutliches Übergewicht literarischer Texte im Vergleich zu journalistischen (26 vs. 2); umgekehrt gibt es bei den „späteren" Texten ein Übergewicht journalistischer im Vergleich zu den literarischen (23 vs. 7).

Tab. 4: Heterogenität der Textkorpus-Zusammensetzung

	K	KJ	LP	JT	< 1920	≥ 1920	Σ
K	❑	❑	17	25	12	30	42
KJ	❑	❑	16	0	16	0	16
LP	17	16	❑	❑	26	7	33
JT	25	0	❑	❑	2	23	25
< 1920	12	16	26	2	❑	❑	28
≥ 1920	30	0	7	23	❑	❑	30
Σ	42	16	33	25	28	30	58

Das Ungleichgewicht in der textuellen Zusammensetzung unserer Stichprobe wird folglich auch deutlich, wenn man (unter Vernachlässigung der Tatsache, daß es sich sowohl beim 'Typ' als auch beim 'Autor' um nominal kodierte Variablen handelt) Korrelationsanalysen durchführt, die in Form von einfachen Haupteffekten den Zusammenhang zwischen den Variablen deutlich machen (vgl. Tab. 5).

Tab. 5: Zusammenhänge zwischen möglichen Einflußfaktoren auf die Wortlänge

	Typ	Jahr	Autor
Typ	_ _ _	$r = 0{,}70$ $p < 0{,}001$	$r = 0{,}54$ $p < 0{,}001$
Jahr	$r = 0{,}70$ $p < 0{,}001$	_ _ _	$r = 0{,}64$ $p < 0{,}001$
Autor	$r = 0{,}54$ $p < 0{,}001$	$r = 0{,}64$ $p < 0{,}001$	_ _ _

Mit diesen Beobachtungen sind zwei wesentliche, in der Auswirkung nicht voneinander unabhängige Schlußfolgerungen verbunden:

1. Falls eine (oder mehrere) der genannten Variablen – Ko-Autorschaft, Funktionalstil, Entstehungszeit – Einfluß auf die durchschnittliche Wortlänge haben, ist eine Beurteilung der mittleren Wortlänge allgemein in der Prosa Čapeks auf der Basis unserer Stichprobe unzulässig.

2. Aufgrund der wechselseitigen Beeinflussung der Variablen ist zu erwarten, daß es bei einfachen Mittelwertvergleichen zu Überlagerungen der Effekte kommt, so daß für eine zuverlässige Entscheidung komplexere Verfahren zur Anwendung kommen müssen.

Schauen wir uns mit diesen *caveats* die Auswirkungen dieser Überlegungen beispielhaft an unserem konkreten Datenmaterial an. Stellen wir als erstes, um den Einfluß der Variablen 'Autorschaft' auf die mittlere Wortlänge zu prüfen, die 40 Karel Čapek allein zuzuschreibenden Texte den 16 Texten gegenüber, die er mit seinem Bruder Josef gemeinsam geschrieben hat. Es stellt sich heraus, daß sich die durchschnittliche Wortlänge der beiden Textgrup-

pen bei $\bar{x}_K$ = 2,02 (s = 0,13) vs. $\bar{x}_{JK}$ = 2,03 (s = 0,08) praktisch nicht voneinander unterscheidet, wie auch ein entsprechender *t*-Test für unabhängige Stichproben zeigt ($t_{FG=46}$ = 0,33; p = 0,74). Sobald man jedoch die journalistischen Texte aus der Analyse ausklammert und die Untersuchung auf die verbleibenden 33 literarischen Texte beschränkt, ergibt sich ein vollkommen anderes Bild: In diesem Falle beläuft sich die mittlere Wortlänge der 17 Texte von Karel Čapek auf $\bar{x}_K$ = 1,89 (s = 0,05), die der 16 gemeinsam verfaßten Texte auf $\bar{x}_{JK}$ = 2,03 (s = 0,08) – ein Unterschied, der durch den *t*-Test als hoch signifikant ausgewiesen wird ($t_{FG=25}$ = 6,08; p < 0,001).

Damit sieht es so aus, als ob ein allfälliger, durch die Variable 'Autorschaft' bedingter Unterschied (d.h. ein Unterschied zwischen den allein und den gemeinsam verfaßten Texten) bei gleichzeitiger Einbeziehung auch der journalistischen Texte überdeckt würde. Das würde jedoch bedeuten, daß (auch) die Variable 'Texttyp' Einfluß auf die mittlere Wortlänge haben müßte, und zwar dergestalt, daß die journalistischen Texte sich durch eine insgesamt höhere Durchschnittslänge als die literarischen Texte auszeichnen.

Dies führt zur Frage, ob in der Tat ein solcher Unterschied zwischen den beiden Texttypen „literarische Prosa" (LP) und „journalistische Prosa" (JT) besteht. In der Tat stellt sich die mittlere Wortlänge der 33 literarischen Texten mit $\bar{x}_{LP}$ = 1,95 (s = 0,09) und der 25 journalistischen Texte mit $\bar{x}_{JT}$ = 2,10 (s = 0,10) als hoch signifikant dar ($t_{FG=50}$ = 5,70; p < 0,001). Der Unterschied zwischen beiden Texttypen ist unvermindert signifikant ($t_{FG=37}$ = 9,21; p < 0,001), wenn man die 25 journalistischen Texte nur den 17 literarischen Texten, die von Karel Čapek allein geschrieben wurden (LP-K), gegenüberstellt ($\bar{x}_{LP\text{-}K)}$ = 1,89; s = 0,05); eine vergleichende Analyse mit gemeinsam verfaßten journalistischen Texten ist, wie oben dargestellt, aufgrund fehlender entsprechender Texte nicht möglich.

Somit scheint sich die oben vorgetragene Vermutung zu bestätigen, daß ein nachweislicher Unterschied, der zwischen den allein und den gemeinsam verfaßten literarischen Texten besteht, durch die Variable 'Texttyp' überlagert wird. Allerdings erweist sich ein Vergleich der 16 gemeinsam verfaßten literarischen Texte ($\bar{x}_{JK}$ = 2,03; s = 0,08) mit den 25 journalistischen Texten von Karel Čapek ($\bar{x}_{JT}$ = 2,10; s = 0,10) ebenfalls als hoch signifikant ($t_{FG=38}$ = 2,77; p < 0,01).

Es ergeben sich damit in unserem Datenmaterial aufgrund der mittleren Wortlänge (zumindest) drei zu differenzierende Untergruppen, deren Charakteristika in der Tab. 6 kurz zusammengefaßt sind; neben der jeweiligen Stichprobengröße (N) enthält diese die Mittelwerte ($\bar{x}$), Standardabweichungen (s) und Standardfehler der Mittelwerte ($s_{\bar{x}}$).

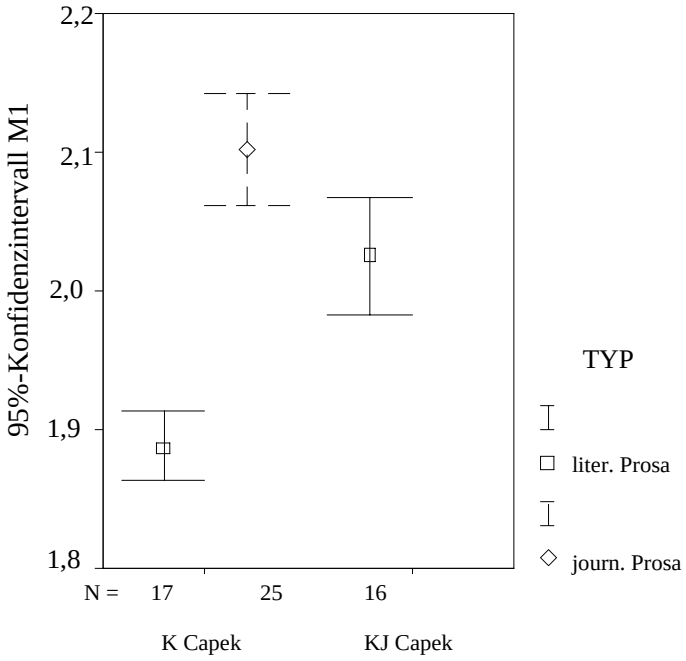
Tab. 6: Unterschiedliche Wortlänge in verschiedenen Text-Untergruppen

Typ	N	$\bar{x}$	s	$s_{\bar{x}}$
LP-KJ	16	2,03	0,08	0,019
LP-K	17	1,89	0,05	0,012
JT-K	25	2,10	0,10	0,020

Wie das Fehlerbalkendiagramm in Abb. 1 zeigt, überschneiden sich die drei 95%-Konfidenzintervalle nur in einem Fall geringfügig.

Die Beobachtung, daß sich innerhalb unseres Datenmaterials ganz offensichtlich allein aufgrund der mittleren Wortlänge unterschiedliche Untergruppen ausgliedern lassen, führt zu der eingangs gestellten Frage zurück, inwiefern sich gegebenenfalls formale und inhaltliche Aspekte der Prosa Čapeks zusammenführen lassen, bzw. inwiefern die von Mukařovský vorgeschlagene Textkategorisierung auch auf formaler Ebene eine Entsprechung finden könnte.

Abb. 1: Fehlerbalkendiagramm dreier Text-Untergruppen



Bevor wir der zuletzt gestellten Frage abschließend detaillierter nachgehen, gilt es jedoch, eine andere Frage nicht zu übersehen, nämlich, ob nicht gegebenenfalls auch chronologische Aspekte der Textentstehung auf die Wortlänge einwirken und gegebenenfalls die beobachteten Unterschiede überlagern.

Bei dem Versuch, einen Zusammenhang zwischen der mittleren Wortlänge in den Texten und der Entstehungszeit der Texte zu überprüfen, bietet es sich an, die entsprechenden Jahreszahlen der Veröffentlichung in eine Rangreihenfolge umzukodieren, diese als 'Jahresrang' zu bezeichnen und die weiteren Berechnungen mit diesen Rangwerten anzustellen; damit ist gleichzeitig die für eine Korrelationsanalyse vorausgesetzte Normalverteilung der Daten gegeben – die bei den originalen Jahreszahlen nicht gegeben ist, wie ein Kolmogorov-Smirnov-Test auf Normalverteilung nach Lilliefors-Korrektur bei einem Wert von 0,123 (p = 0,028) zeigt .

Es stellt sich heraus, daß die Korrelation zwischen der mittleren Wortlänge und der in eine Rangfolge transformierten Jahreszahlen bei einem Spearmanschen Korrelationskoeffizienten von r = .20 kein Signifikanzniveau erreicht (p = .12). Dieser Befund, daß das Entstehungsjahr der jeweiligen Texte

offenbar keinen signifikanten Einfluß auf die mittlere Wortlänge hat, deckt sich mit dem Ergebnis einer Kovarianz-Analyse – zum Verfahren der Varianzanalyse s.u. –, durchgeführt mit der abhängigen Variable 'mittlere Wortlänge', den festen Faktoren 'Autorschaft' und 'Texttyp' sowie der Kovariaten 'Jahresrang'. Denn hierbei stellt sich ein hochsignifikanter Einfluß der beiden Faktoren 'Autorschaft' (F = 12,09; p = 0,001) und 'Texttyp' (F = 51,99; p < 0,001) heraus, der Einfluß der Kovariaten 'Jahresrang' hingegen als nicht signifikant (F = 0,04, p = 0,84). Der Umstand, daß der Einfluß der Kovariaten 'Jahresrang' nicht im mindesten an eines der traditionellen Signifikanzniveaus herankommt, erlaubt es im übrigen, die durch einen Levene-Test nachzuweisende Verletzung der vorausgesetzten Varianzenhomogenität (F = 3,58; p = 0,035) zu vernachlässigen. An dem dargestellten Gesamtergebnis ändert sich in seiner Deutlichkeit übrigens auch bei einer Berechnung über die originalen Jahresdaten nichts.

Die soeben getroffene Feststellung schließt allerdings die Möglichkeit aus, daß es dennoch einen Einfluß der Entstehungszeit auf die mittlere Wortlängen geben könnte, und zwar dann, wenn man nicht die Auswirkung auf das gesamte Textkorpus überprüft, sondern auf Teilstichproben. Diese Überprüfung scheint insofern sinnvoll zu sein, als auch bei der Überprüfung eines möglichen Einflusses der Entstehungszeit der Texte auf die mittlere Wortlänge allfällige Überlagerungen von Effekten nicht übersehen werden. Deshalb ist es angebracht, sowohl für die beiden Texttypen als auch für die gemeinsam und allein verfaßten Texte getrennte Korrelationsberechnungen anzustellen. Die Ergebnisse sind der Tab. 7 zu entnehmen.

Tab. 7: Korrelationen zwischen Wortlänge und Entstehungszeit in verschiedenen Textgruppen

	r	*p*
LP	.74	< 0,001
JT	.16	.45
LP-K	.42	< 0,01
LP-JK	.29	.27

Wie zu sehen ist, unterliegt die journalistische Prosa im Gegensatz zur literarischen Prosa offenbar keinerlei chronologischen Einflüssen; zudem sind es bei der literarischen Prosa nur die von Karel Čapek verfaßten Texte, bei denen sich die Entstehungszeit auf die mittlere Wortlänge auswirkt, während die gemeinsam verfaßten Texte, was die Wortlänge angeht, eine von der Entstehungszeit relativ unabhängige Textgruppe zu bilden scheinen.

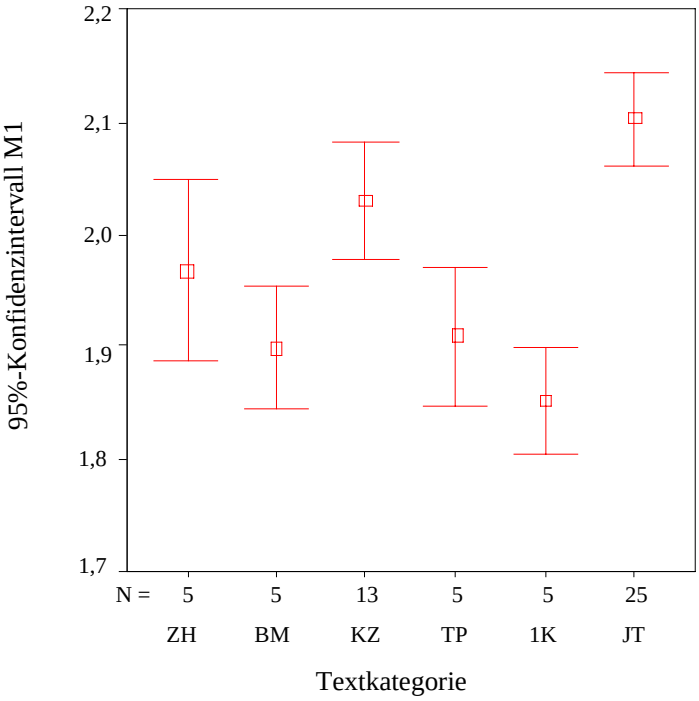
Damit kommen wir, um zu einer ersten Zusammenfassung zu gelangen, aufgrund unserer exemplarischen Analysen zu der vorläufigen Schlußfolgerung, daß die Wortlänge offenbar durch verschiedene Faktoren beeinflußt werden kann, daß sich jedoch der jeweilige Einfluß eines oder mehrerer dieser (potentiellen) Einflußfaktoren weder auf die gesamte Stichprobe noch in identischer Art und Weise auf zu unterscheidende Teilstichproben auswirkt, wohl aber die Bildung verschiedener Untergruppen zur Folge hat.

Diese Option der Existenz möglicher Untergruppen führt schließlich zurück zu der eingangs aufgeworfenen Frage, inwiefern sich eine Unterglie-

derung Čapekscher Texten à la Mukařovský auch auf formaler Ebene spiegelt, bzw. inwiefern eine solche Spiegelung transparent gemacht werden kann. Unternehmen wir also abschließend einen Versuch, die Existenz unterschiedlicher Untergruppen nachzuweisen, ohne dabei die Frage danach zu stellen, welche Faktoren für die Distinktion allfällig existierender Unterschiede verantwortlich sind. Geben wir zu diesem Zweck die relativ grobe Untergliederung der drei Texttypen in *LP* und *JT* auf, und vernachlässigen wir auch die Frage der individuellen vs. kollektiven Autorschaft. Ordnen wir statt dessen jedem einzelnen Text als kategoriales Merkmal den Band zu, in dem er jeweils erschienen ist (vgl. Tab. 4); die journalistischen Texte erhalten analog dazu *in toto* das Merkmal '*JT*'. Insgesamt ergeben sich somit sechs Kategorien: *ZH, BM, KZ, TP, 1K, JT.*

Führen wir mit diesen Angaben eine einfaktorielle Varianzanalyse durch; diese dient dazu, mehr als zwei unabhängige Stichproben hinsichtlich ihrer Mittelwerte miteinander zu vergleichen. Ungeachtet der irreführenden Bezeichnung 'Varianzanalyse' geht es dabei freilich nicht um die Überprüfung der Varianzen auf signifikante Unterschiede, sondern um eine Zerlegung der Gesamt-Varianz. Die Voraussetzungen zur Durchführung der Varianzanalyse sind als gegeben anzusehen, da (a) bei derart geringen Fallzahlen eine signifikante Abweichung von der Normalverteilung in den einzelnen Stichproben kaum möglich ist, wenn nicht eklatante Ausreißer in den Werten auftreten, und (b) die Varianzenhomogenität gewährleistet ist, wie ein Levene-Test zeigt (F = 1,95; p = 0,10). Es stellt sich heraus, daß in der Tat hinsichtlich der mittleren Wortlänge ein hoch signifikanter Unterschied zwischen den sechs unterschiedenen Kategorien besteht ($F_{FG=5}$ = 13,04; p < 0,001). Die Unterschiede sind in der Abb. 2 veranschaulicht.

Abb. 2: Fehlerbalkendiagramm der sechs Text-Untergruppen



Auf den ersten Blick stechen die journalistischen Texte (*JT*) sowie die Texte aus den *Povídky z jedné kapsy* (*1K*) ins Auge, auch scheint eine deutlich Nähe zwischen den Texten der Gruppen '*BM*' und '*TP* ' gegeben zu sein. Jegliche weitere Einschätzung der übrigen Textgruppen entzieht sich jedoch einer intuitiven Bewertung 'auf den ersten Blick'. Mit anderen Worten: Ungeachtet der wichtigen Feststellung, daß es signifikante Wortlängen-Unterschiede zwischen den einzelnen Textgruppen gibt, ist damit noch nicht geklärt, welche der Gruppenmittelwerte sich im einzelnen paarweise voneinander unterscheiden. Diese Frage läßt sich jedoch mit sogenannten Post-Hoc-Mittelwert-Vergleichen klären, von denen es mittlerweile mehr als ein Dutzend verschiedener gibt. Einer der am häufigsten verwendeten ist der sog. Scheffé-Test, mit dem gemeinsame paarweise Vergleiche gleichzeitig für alle möglichen paarweisen Kombinationen der Mittelwerte durchgeführt werden. Dieser Test wird als relativ robust angesehen (weil er gegenüber Verletzungen von Voraussetzungen unempfindlich ist), und er gilt als konservativ (weil Mittelwertunterschiede erst bei relativ großen Differenzen als gesichert angesehen werden). Wenden wir also den Scheffé-Post-Hoc-Test auf unser Datenmaterial an; die Ergebnisse sind der Tab. 8 zu entnehmen.

Tab 8: Ergebnisse der Post-Hoc-Mittelwert-Vergleiche (*Scheffé-Test*)

Band	*N*	Untergruppe für α = .05		
		1	2	3
Povídky z jedné kapsy	5	1,8508		
Boží Muka	5	1,8979	1,8979	
Trapné povídky	5	1,9088	1,9088	
Zářivé hlubiny	5	1,9676	1,9676	1,9676
Krakonošova zahrada	13		2,0294	2,0294
Journalistik	25			2,1026
Signifikanz		0,298	0,180	0,158

Deutlich erkennbar lassen sich aufgrund des Scheffé-Tests vier homogene Untergruppen untergliedern: Ohne jede Frage stellen die journalistischen Texte und die Texte aus den *Povídky z jedné kapsy* jeweils eigene Untergruppen dar; ebenso deutlich bilden die Erzählungen aus *Boží Muka* und den *Trapné povídky* eine gemeinsame Untergruppe. Weiterhin heben sich die frühen Erzählungen aus *Krakonošova zahrada* deutlich von den übrigen Textgruppen ab. Lediglich die Erzählungen aus *Zářivé hlubiny* nehmen in gewissem Sinne eine Zwischenstellung ein: Sie lassen sich einerseits den (frühen) Texten aus *Krakonošova zahrada* zuordnen (wie auch Mukařovský das bei seiner Periodisierung tat); andererseits stellen sie ohne Frage einen Übergang von den früheren Texten zu den (späteren) Texten der Gruppe *Boží Muka* und *Trapné povídky.* In diesem Zusammenhang sei daran erinnert, daß es sich gerade bei den Texten aus *Zářivé hlubiny* um eine Mischung von Texten handelt, die zum Teil bereits von Karel Čapek allein, zum Teil aber auch noch mit seinem Bruder gemeinsam verfaßt wurden.

Auf jeden Fall bestätigt sich somit unsere allgemeine Annahme, daß sich allein aufgrund des formalen Kriteriums der Wortlänge eine Untergliederung

der Texte in homogene Untergruppen vornehmen läßt. Daß darüber hinaus die sich ergebenden Untergruppen mit den von Mukařovský vor fast 70 Jahren aus ganz anderen Motiven besprochenen Textgruppen konvergieren, kann dabei in der Tat nicht nur als starke Evidenz für die dialektische Wechselbeziehung inhaltlicher und formaler Aspekte künstlerischer Texte, sondern allgemein als Nachweis der Produktivität quantifizierender Verfahren im Bereich der Text- und Kulturwissenschaften allgemein angesehen werden.

Literatur

Doležel, L. (1965): Zur statistischen Theorie der Dichtersprache. In: *Kreuzer/Gunzenhäuser* (Hg.) (1965), 275–293.

Doležel, L. (1967): The Prague School and the Statistical Theory of Poetical Language. *Prague Studies in Mathematical Linguistics 2*, 97–104.

Gendolla, P., Kamphusmann, Th. (Hgg.) (1999): *Die Künste des Zufalls*. Frankfurt/M.

Grotjahn, R., Altmann, G. (1993): Modelling the Distribution of Word Length: Some Methodological Problems. In: Köhler, R., Rieger, B. (eds.), *Contributions to Quantitative Linguistics*, Dordrecht etc., 141–153.

Kreuzer, H., Gunzenhäuser, R. (Hgg.) (1965): *Mathematik und Dichtung*. München (^{4}1971).

Ludvíková, M. (1972): Quantitative Syllable Analysis of Words in Czech. *Prague Studies in Mathematical Linguistics 3*, 27–34.

Mukařovský, J. (1934) (ed.): *Výbor z prózy Karla Čapka*. Praha (^{2}1946).

Mukařovský, J. (1934): Die Entwicklung von K. Čapeks Prosa. In: Ders., *Kapitel aus der Poetik*, Frankfurt/M. 1967, 55–107 [= Vývoj Čapkovy prózy].

Uhlířová, L. (1995): On the Generality of Statistical Laws and Individuality of Texts. A Case of Syllables, Word Forms, their Length and Frequencies. *Journal of Quantitative Linguistics* 2/3, 238–247.

Uhlířová, L. (1996): How Long Are Words in Czech? In: Schmidt, P. (ed.), *Issues in General Linguistic Theory and The Theory of Word Length* (Glottometrika 15), Trier, 134–146.

Wimmer, G., Altmann, G. (1996): The Theory of Word Length Distribution: Some Results and Generalizations. In: Schmidt, P. (ed.), *Issues in General Linguistic Theory and The Theory of Word Length* (Glottometrika 15), Trier, 112–133.

Wimmer, G., Köhler, R., Grotjahn, R., Altmann, G. (1994): Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics* 1, 98–106.

Wimmer, G., Witkovský, V., Altmann, G. (1999): Modification of Probability Distributions. Applied to Word Length Research. *Journal of Quantitative Linguistics* 6/3, 257–268.

❦

Anhang I: Verzeichnis der analysierten literarischen Texte

Autor	Text	Nr.	Jahr	Bibl.	Band	N	$\bar{x}$	s
JK	*Alkohol*	1	1909	75	KZ	544	2,0533	1,0105
JK	*Aristokracie*	2	1909	58	KZ	652	2,0245	1,0373
JK	*Čas*	3	1909	88	KZ	426	2,0047	1,0963
JK	*Večeře*	4	1909	92	KZ	414	2,1667	1,0807
JK	*O smrti a zradě života*	5	1909	71	KZ	484	1,9375	0,9923
JK	*Zářivé hlubiny*	6	1913	192	ZH	2708	1,9952	1,0423
K	*Mezi dvěma polibky*	7	1911	169	ZH	1519	1,9454	1,0281
K	*Ostrov*	8	1912	174	ZH	2496	1,8690	0,9834
JK	*L'éventail*	9	1910	139	ZH	3849	1,9891	1,0666
JK	*Červená povídka*	10	1910	134	ZH	3962	2,0389	1,0620
K	*Utkvění Času*	11	1913	230	BM	457	1,9212	0,9958
K	*Šlépěj*	12	1916	259	BM	1608	1,9160	0,9886
K	*Zatracená cesta*	13	1916	268	BM	2373	1,8310	0,9355
K	*Historie beze slov*	14	1917	3155	BM	875	1,9451	1,0116
K	*Pokušení*	15	1917	3155	BM	984	1,8760	0,9178
K	*Tři*	16	1919	479	TP	2082	1,9265	0,9936
K	*Helena*	17	1919	476	TP	3288	1,9361	1,0285
K	*Surovec*	18	1921	651	TP	4327	1,9392	0,9843
K	*Košile*	19	1920	642	TP	2327	1,9252	0,9792
K	*Uražený*	20	1919	514	TP	3852	1,8214	0,9325
K	*Modrá chryzantéma*	21	1928	1749	1K	1740	1,8569	0,9545
K	*Rekord*	22	1928	1769	1K	2147	1,8039	0,9024
K	*Šlépěje*	23	1928	1760	1K	2335	1,9071	0,9771
K	*Kupón*	24	1928	1771	1K	2558	1,8331	0,9605
K	*Případy pana Janíka*	25	1928	1785	1K	2525	1,8555	1,0168
JK	*Antika a křesťanství*	26	1909	77	KZ	582	2,1495	1,0988
JK	*Morálka (I)*	27	1909	84	KZ	401	2,0050	1,0161
JK	*Morálka (2)*	28	1909	86	KZ	351	1,8860	0,9629
JK	*Benátský karneval*	30	1910	137	KZ	787	2,0928	1,1329
JK	*O různých prostředcích*	31	1912	189	KZ	396	1,9419	0,9725
JK	*Jarní improvizace*	32	1909	64	KZ	725	2,0924	1,1085
JK	*Ranní*	33	1909	94	KZ	339	2,0767	1,0362
JK	*Fejeton*	34	1910	159	KZ	1952	1,9508	1,0178

In der vorstehenden Tabelle finden sind in den einzelnen Spalten:

1. Autorschaft (JK = Josef + Karel Čapek; K = Karel Čapek)
2. Titel des Textes
3. Nummer innerhalb des analysierten Textkorpus
4. Erscheinungsjahr
5. Verweis gem.: Mědílek, Boris et al. (eds.): *Bibliografie Karla Čapka*. Praha 1990.
6. Sammelband (vgl. Tab. 4)
7. Anzahl der Worte im gegebenen Text (N)
8. Mittlere Wortlänge im gegebenen Texte ($\bar{x}$)
9. Standardabweichung der Wortlänge im gegebenen Texte (s)

Anhang II: Verzeichnis der analysierten journalistischen Texte

Text	Nr.	Jahr	Publiziert	N	$\bar{x}$	s
Noviny a věda	1	1917	*Lidové Noviny, 28.10.*	857	2,0175	1,1333
Román o těle	2	1919	*Lidové Noviny, 23.12.*	1116	2,0923	1,0783
Místo kritiky	3	1920	*Lidové Noviny, 24.12.*	976	2,1793	1,2396
Příklad moskevských	4	1921	*Lidové Noviny, 23.5.*	884	2,0735	1,091
O jedné obrazárně	5	1922	*Lidové Noviny, 18.3.*	591	2,3215	1,1554
Posedlost	6	1922	*Lidové Noviny, 17.5.*	591	1,8917	0,9479
Nemohu mlčet	7	1922	*Lidové Noviny, 10.6.*	882	1,9887	1,0879
5000	8	1923	*Lidové Noviny, 13.1.*	541	2,1257	1,1141
Za malou Prahu	9	1924	*Lidové Noviny, 05.11.*	914	2,1783	1,2207
Kterak se čtou knihy	10	1925	*Lidové Noviny, 13.12.*	624	2,0465	1,0754
O tom nacionalismu	11	1926	*Lidové Noviny, 25.07.*	761	2,0999	1,2058
Žurnalistům	12	1926	*Lidové Noviny, 5.9.*	949	2,0021	1,0638
Otázka divadelní	13	1926	*Lidové Noviny, 17.12.*	997	2,2307	1,1441
Do mrtvé sezóny	14	1927	*Lidové Noviny, 3.7.*	789	2,0697	1,0817
Z žiněného hrobu	15	1928	*Lidové Noviny, 17.8.*	475	1,9263	0,9673
Kantoři a škola	16	1929	*Lidové Noviny, 28.04.*	616	2,1834	1,2537
Knížky životní moudrosti	17	1930	*Almanach Kmene*	341	2,1349	1,0609
Pražký román Karla Poláčka	18	1931	*Lidové Noviny, 11.02.*	524	2,1947	1,1853
Tíseň	19	1932	*Lidové Noviny, 23.12.*	664	1,9578	1,0899
O dnešním jazyce	20	1932	*Lidové Noviny, 7.2.*	724	2,0704	1,1666
O detektivkách	21	1932	*Lidové Noviny, 14.2.*	443	2,1986	1,2024
Evropa	22	1934	*Život*	724	2,1354	1,2104
Volební advent	23	1935	*Lidové Noviny, 1.5.*	1069	2,1169	1,1720
Varovné znamení	24	1937	*Čin, 14.1.*	537	2,162	1,1343
Kam směřuje vývoj	25	1938	*Přítomnost, 5.1.*	919	2,1665	1,1576