Peter Grzybek (Graz)

Randbemerkungen zur Korrelation von Wort- und Silbenlänge im Kroatischen

> Once a correlation has been found it must, then, be checked by further experiment and/or theory. But, again, a checked correlation should not be the goal of research but the basis for posing a further problem, namely: What is the *mechanism* underlying the correlation, i.e. what brings it about and in accordance with what laws?
>
> **Mario Bunge (1967: 277)**

0. Fragestellung

Der Korrelationsbegriff unterliegt sowohl in der Alltags- als auch in der Wissenschaftssprache nicht nur unterschiedlichen Verstehensweisen, sondern er wird häufig unscharf und zum Teil inflationär verwendet. Im vorliegenden Beitrag soll die Tragfähigkeit eines ***operationalen*** bzw. ***operationalisierbaren Korrelationsbegriffs*** veranschaulicht werden, der sich der statistischen Methodenlehre verdankt. Aus Raumgründen muß dabei auf tiefergehende oder weiterführende Analysen verzichtet werden; statt dessen soll an einem ausgewählten Beispiel die Anwendungsmöglichkeit dieses Korrelationsbegriffs demonstriert werden.

1. Der Korrelationsbegriff

Probleme der Verwendung des Korrelationsbegriffs betreffen (zumindest) zweierlei Aspekte:

1. zum einen die Frage danach, welche Entitäten als miteinander korreliert angesehen werden – denn wenn die Verwendung des Korrelationsbegriffs, wie es nicht selten der Fall ist, darauf hinaus läuft, daß in letzter Konsequenz „alles mit allem“ korreliert, wird der Korrelationsbegriff als solcher unsinnig, weil er keine Erklärungskraft mehr hat;
2. zum anderen die konkrete Definition des Korrelationsbegriffs und seiner Abgrenzung zu verwandten Begriffen, die notwendige Voraussetzung dafür ist, daß der Begriff der Korrelation eine spezifische Erklärungskraft beinhaltet.

In erster Linie ist der Begriff der Korrelation von dem (noch) allgemeineren und umfassenderen Begriff der **Relation** abzugrenzen, wobei unter einer Relation eine **spezifische dyadische Beziehung** (eine Beziehung zwischen zwei Elementen also) verstanden werden soll. Bekannte spezifisch mathematische Relationen sind z.B. die in Tab. 1. aufgeführten:

Tab. 1: Relationen

Beziehung		Bedeutung
a = b		a ist gleich b
a < b		a ist kleiner als b
a > b		a ist größer als b
a ≤ b		a ist gleich oder kleiner als b
a ≥ b		a ist gleich oder größer als b
a ≅ b	}	a ist ungefähr gleich b,
a ≈ b	}	a ist angenähert b
a ≠ b		a ist nicht gleich b

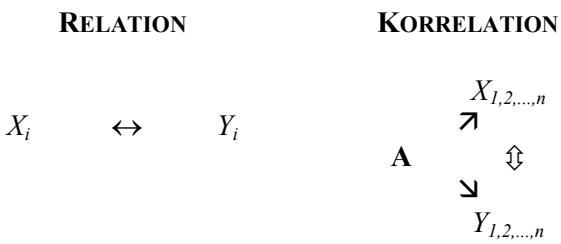
Comment: = I (1510-1558)

Comment: <, > (1560-1621)

Comment: <, > (1560-1621)

Im Gegensatz dazu betrifft eine **Korrelation** die **Abhängigkeit zwischen den Ausprägungen zweier Merkmale (Variablen)** ***X,Y*** eines Individuums, Prozesses o.ä. Der Unterschied zwischen Relation und Korrelation läßt sich demnach wie in Abb. 1 veranschaulichen.

Abb. 1: Relation und Korrelation



Vor dem Hintergrund dieser Differenzierung läßt sich eine **Typisierung korrelativer Zusammenhänge** erstellen (vgl. Sachs 1992: 507ff.), die auf die verschiedenen möglichen Bedingungen einer Korrelation abhebt. So kann eine Korrelation bedingt sein durch: a) rein formale Faktoren, b) Heterogenität des Materials ("Inhomogenitätskorrelation"), c) gemeinsame Abhängigkeit von dritten Größen, d) kausale Faktoren.

1.1. *Formale Korrelationen* (a), die oft fälschlicherweise auch als „Scheinkorrelationen" bezeichnet werden, sind nicht auf dritte Einflußgrößen zurückzuführen, hängen aber auch nicht kausal miteinander zusammen.– Wenn es sich bei den Variablen *X* und *Y* beispielsweise um zwei sich miteinander zu 100% ergänzende Prozentsätze handelt, muß zwangsläufig zwischen ihnen eine negative Korrelation vorliegen.

1.2. Die *Inhomogenitätskorrelation* (b) ist, wie auch die Gemeinsamkeitskorrelation (c), durch den Einfluß von Drittvariablen bedingt.

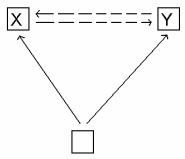
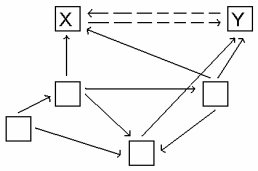
Bei der *Inhomogenitätskorrelation* steht im Vordergrund, daß das Material aus verschiedenen Teilmassen besteht, die in verschiedenen Bereichen des Koordinatensystems liegen. – Ein Beispiel wäre ein positiver Zusammenhang von Schuhgröße und Einkommen, der dadurch zustande kommt, daß man eine Gesamtstichprobe von Frauen und Männern betrachtet (ohne daß dieser Zusammenhang bei einer der Teilstichproben der Frauen bzw. Männer zu beobachten ist), wobei Frauen, die üblicherweise kleinere Schuhgrößen haben, aus ganz anderen Gründen geringere Einkommen haben als Männer.

1.3. Bei der *Gemeinsamkeitskorrelation* (c) kommt eine beobachtete Abhängigkeit durch den Einfluß einer dritten Größe zustande. – Ein Beispiel wäre die Länge des linken und des rechten Arms einer Person, ein anderes bekanntes Beispiel ist ein positiver Zusammenhang zwischen der abnehmenden Zahl an Storchennestern und abnehmender Geburtenrate in einer Region, der z.B. auf zunehmende Industrialisierung in der Region zurückzuführen ist.

1.4. Eine *kausale Korrelation* (d) basiert auf einer *direkten Abhängigkeit* des Merkmals *Y* vom Merkmal *X*.– Beispiel wären hier der Zusammenhang zwischen Begabung und Testleistung, zwischen Arbeitszeit und Preis o.ä.

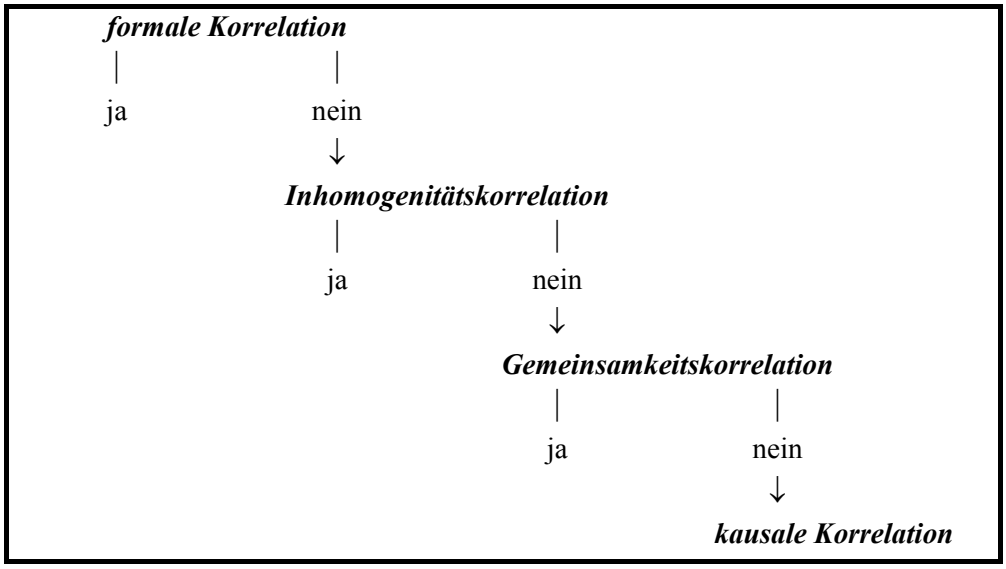
Eine beobachtete Korrelation kann also sehr unterschiedliche Ursachen haben, die sich in der Art ihrer Wirksamkeit wie in den Abb. 3a-e nach Diehl/Kohr (1979: 205) veranschaulichen lassen:

Abb. 3: Korrelationen

(a)	X ⟶ Y	Der Zusammenhang kommt durch den Einfluß von *X* auf *Y* zustande.
(b)	X ⟵ Y	Der Zusammenhang kommt durch den Einfluß von *Y* auf *X* zustande.
(c)	X ⟷ Y	Der Zusammenhang kommt durch den Einfluß von *X* auf *Y* und von *Y* auf *Y* zustande.
(d)		Ein dritter Faktor beeinflußt sowohl *X* als auch *Y* und führt zu dem zwischen *X* und *Y* festgestellten Zusammenhang; dies schließt einen (zusätzlichen) Einfluß von *X* auf *Y* und/oder *Y* auf *X* nicht aus.
(e)		Ein Bündel von miteinander in Beziehung stehenden Variablen beeinflußt sowohl *X* als auch *Y* und führt zu dem zwischen *X* und *Y* festgestellten Zusammenhang; dies schließt einen (zusätzlichen) Einfluß von *X* auf *Y* und/oder *Y* auf *X* nicht aus.

Die Anerkennung einer *kausalen Korrelation* erfolgt de facto durch Ausschluß der anderen Möglichkeiten (vgl. Abb. 2). Von grundsätzlicher Bedeutung ist in diesem Zusammenhang, daß zwar die Feststellung von Zusammenhängen Gegenstand der Statistik ist, daß aber die Prüfung eines gefundenen sachlichen Zusammenhangs auf mögliche *kausale* Zusammenhänge außerhalb der statistischen Methodenlehre liegt. Mit anderen Worten: Das Vorliegen einer Korrelation zwischen zwei Variablen *X* und *Y* impliziert nicht notwendigerweise eine *kausale* Beziehung zwischen ihnen: Wenn *X* und *Y* kovariieren, so ist dies eine notwendige, aber keine hinreichende Bedingung für eine Aussage über eine kausale Beziehung zwischen ihnen.

Abb. 2: Typologie der Korrelation



Eine ***Korrelationsanalyse*** ermittelt den Grad (die Stärke) und die Art des Zusammenhangs zwischen den Merkmalen *(X,Y)*. Im Falle eines **linearen Zusammenhangs** ist der Pearson'sche Korrelationskoeffizient r ($-1 \le r \le +1$) eine Maßzahl für Stärke und Richtung des Zusammenhangs: Für $r = 0$ besteht kein linearer Zusammenhang, für $r = \pm 1$ besteht ein funktionaler (positiver oder negativer) Zusammenhang. Durch eine ***Regressionsanalyse*** werden die beobachteten Werte einer Regressionsgleichung angepaßt, aufgrund derer sich die abhängige Variable schätzen (vorhersagen) läßt. Im Falle einer linearen Regression lautet die Regressionsgleichung:

(1) $y = a + bx$.

Hierbei ist a eine Konstante, welche die Höhe der Regressionsgeraden (Schnittpunkt mit der y-Achse) angibt, und b ist der Regressionskoeffizient, der die Steilheit des Anstiegs bzw. Abfalls der Geraden bestimmt. Im Falle einer **nicht-linearen (bzw.**

kurvilinearen) Regression ergibt das Verhältnis von *X* und *Y* eine bestimmte Kurve, die sich an Regressionsgleichungen der folgenden Art anpassen läßt, wobei die Grundidee darin besteht, die Parameter der Gleichung, das heißt die Variablen und Konstanten, so zu berechnen, daß die Abweichungen zwischen den empirischen und den theoretischen Werten minimal werden. Tab. 2 führt einige Beispiele bekannter und häufig verwendeter Regressionsgleichungen an.

Tab. 2: Regressionsgleichungen

Inverse Regression	$y = a + b/x$
Quadratische Regression	$y = a + bx + cx^2$
Potenzregression	$y = ax^b$
Exponentielle Regression	$y = ae^{bx}$

2. Zur Wortlänge im Kroatischen

Im folgenden sollen die obigen theoretischen Überlegungen an einem konkreten Beispiel nachvollzogen werden. Als Fragestellung soll der Zusammenhang von Wort- und Silbenlänge dienen. Dabei werden wir uns des weiteren in zweierlei Hinsicht einschränken: erstens werden wir uns ausschließlich auf kroatisches Material beschränken, und zweitens auf Wörterbuchmaterial, das sich im Hinblick auf seine Struktur von Texten dadurch unterscheidet, daß jede Einheit nur ein einziges Mal vorkommt. Diese Feststellung ist zwar trivial, für die Berechnung des Zusammenhangs allerdings von zentraler Bedeutung, ohne daß wir hier auf den Zusammenhang näher eingehen könnten.

Bei der aufgewiesenen Frage ist es geboten ist, sich an die Dissertation von Gajić aus dem Jahre 1950 zu erinnern: Gajić hat die durchschnittliche, in Lauten berechnete Silbenlänge aller Einträge in Junckers *Deutsch-Kroatischem Wörterbuch* von 1930 berechnet, und die von ihm angestellten Berechnungen sollen uns im weiteren als Demonstrationsmaterial dienen.

Aufgrund der bei Gajić angegebenen Rohdaten ergeben sich die in der dritten Spalte der Tab. 3 präsentierten Werte.

Tab. 3: Abhängigkeit von Wort- und Silbenlänge nach Gajić (1950)

Silben pro Wort	**Anzahl d. Wörter mit *f* Silben pro Wort**	**Laute pro Silbe (beoachtet)**	**Laute pro Silbe $y = a + bx$**
f	n	ξ	ξ'
1	717	3,45	3,03
2	4038	2,66	2,83
3	6060	2,38	2,62
4	5066	2,20	2,41
5	1239	2,11	2,20
6	145	2,06	1,99
7	14	2,00	1,78

Deutlich kommt die folgende Tendenz zum Ausdruck: Je mehr Silben die Wörter aufweisen, um so kürzer werden im Durchschnitt die Silben. Der Bonner Romanist Paul Menzerath, bei dem die Dissertation von Gajić seinerzeit geschrieben wurde, hat diese Tendenz in seiner *Architektonik des deutschen Wortschatzes* (1954: 101) in eine Verallgemeinerung überführt: „Die relative Lautzahl nimmt bei steigender Silbenzahl ab, oder mit anderer Formel gesagt: je mehr Silben ein Wort hat, um so (relativ) kürzer ist es."

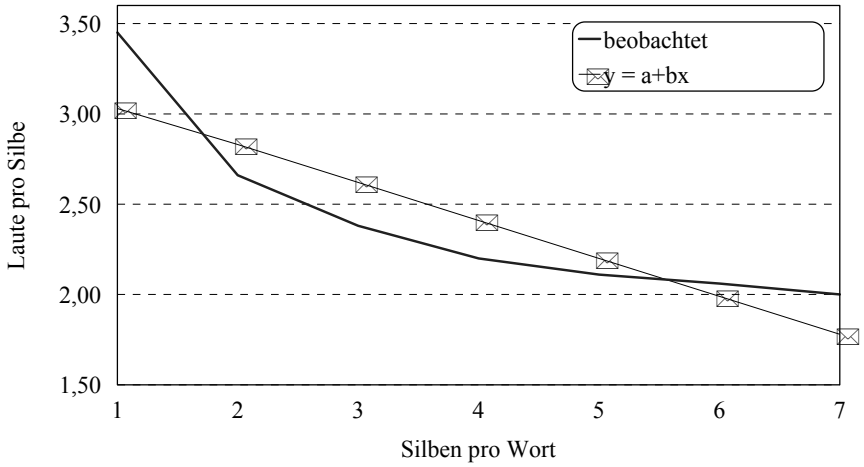
Altmann (1980), der diese Aussage später in den Status des sog. Menzerath'schen Gesetzes erhoben hat, erweiterte seinerseits dieses Gesetz auf alle Sprachebenen, so daß es in der verallgemeinerten Form wie folgt lautet: »Je größer bzw. komplexer ein sprachliches Konstrukt ist, um so kleiner bzw. einfacher sind seine Konstituenten.« Darüber hinaus hat Altmann eine mathematische Formalisierung dieser Tendenz in Form einer nicht-linearen Regressionsgleichung vorgenommen, auf die unten noch einzugehen sein wird; zum Zwecke der Veranschaulichung sei jedoch zunächst das Ergebnis einer linearen Regressionsanalyse präsentiert.

Die Berechnung der Stärke der zu vermutenden Korrelation in Form des Pearson'schen Korrelationskoeffizienten ergibt einen Wert von $r = -.879$; diese negative Korrelation (je mehr Silben pro Wort, desto weniger Laute pro Silbe) ist bei zweiseitiger Prüfung hoch signifikant ($p < 0.01$). Nach Einsetzung der Werte für $a = 3.24$ und $b = -.21$ ergeben sich aufgrund der Regressionsgleichung

(1a) $y = 3.24 - 0.21x$,

die in der vierten Spalte der Tab. 3 (s.o.) erkenntlichen theoretischen Vorhersagewerte, die in der Abb. 3 auch graphisch dargestellt sind.

Abb. 4: Lineares Anpassungsmodell



Es zeigt sich deutlich, daß eine lineare Anpassung zwar möglich, aber nur bedingt überzeugend ist; dies führt uns zurück zu dem oben bereits angesprochenen Menzerath'schen Gesetz und seiner Formalisierung nach Altmann. In seiner allgemeinsten Form (III) lautet die Regressionsgleichung:

(2) $y = ax^b e^{-cx}$

mit den beiden Spezialfällen (I) für $b = 0$, $c \neq 0$ und (II) für $b \neq 0$, $c = 0$.

Tab. 4: Das Menzerath'sche Gesetz in der Formalisierung von Altmann (1980)

I	b = 0		$y = Ae^{-cx}$
II	b ≠ 0	c = 0	$y = Ax^b$
III		c ≠ 0	$y = Ax^b e^{-cx}$

Wie zu erwarten ist, stellt die komplexeste Formel (III) die beste Anpassung dar. Die aufgrund der Regressionsgleichung $y=3.27x^{-0.43}e^{0.05x}$ erhaltenen Ergebnisse sind der vierten Spalten der Tab. 5 zu entnehmen. Die Güte der Anpassung wird durch den sogenannten Determinationskoeffizienten R^2 bestätigt, ein Maß für den Zusammenhang zwischen den beobachteten und den theoretischen Werten im Intervall zwischen 0 und 1. In unserem Fall liegt R^2 nur minimal unter dem Maximum von 1, nämlich bei 0.999 – das heißt, daß die durchschnittliche Silbenlänge zu mehr als 99% durch die Wortlänge determiniert ist. Auf eine graphische Darstellung können wir aus diesem Grunde verzichten.

Allerdings ist Formel (III) auch diejenige mit den meisten Parametern (a,b,c) – in der Regel ist man jedoch bestrebt, solche Anpassungen als gelungener zu betrachten, bei denen die theoretische Schätzung auf eine möglichst geringe Anzahl von Parametern reduziert werden kann. In diesem Sinne wird Formel (II)

üblicherweise als „Standardfall“ für das Menzerath’sche Gesetz angesehen, und mit dieser Formel hat auch Altmann (1989: 55) die Daten von Gajić angepaßt. In der Tat erhalten wir hervorragende Anpassungsergebnisse ($R^2 = .98$), die der fünften Spalte der Tab. 5 zu entnehmen sind. Die entsprechende Regressionsgleichung lautet:

$y = 3.23x^{-0.278}$.

Bis an diesen Punkt sehen wir das Menzerath’sche Gesetz in der von Altmann (1980) vorgeschlagenen Formalisierung als weitestgehend bestätigt. Interessant ist insofern, daß auch noch eine ganz andere Regressionsgleichung eine vergleichbar gute, ja sogar geringfügig bessere Anpassung erzielt, nämlich die von SPSS standardmäßig angebotene Gleichung:

(3) $y = e^{(a+bx)}$

Mit dieser Gleichung kommt die Anpassung nach Einsetzung der Parameter $a = 0.628$ und $b = 0.630$ auf ein R^2 von .985. Nach der von Wimmer / Köhler / Altmann (Ms.) vorgeschlagenen Umformulierung dieser Gleichung in:

(4) $y = Ce^{(-a/x)}$

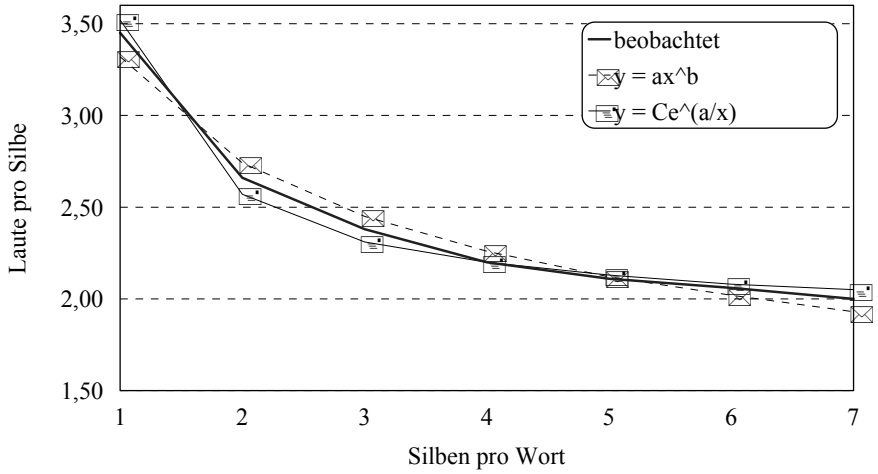
beträgt der Determinationskoeffizient mit den eingesetzten Parametern $C = 1.889$ und $a = 0.613$ in diesem Fall $R^2 = .987$. Die Ergebnisse sind der sechsten Spalte der Tab. 5 zu entnehmen.

Tab. 5: Theoretische Anpassungen der Wort- und Silbenlänge

Silben pro Wort	**Anzahl d. Wörter mit *f* Silben pro Wort**	**Laute pro Silbe (beoachtet)**	**Laute pro Silbe (theoret.)**	**Laute pro Silbe (theoret.)**	**Laute pro Silbe (theoret.)**
f	***n***	***ξ***	***ξ э*** ***($y = ax^b e^{(-cx)}$)***	***ξ э*** ***($y = ax^b$)***	***ξ э*** ***($y = Ce^{(-a/x)}$)***
1	717	3,45	3,44	3,32	3,49
2	4038	2,66	2,68	2,74	2,57
3	6060	2,38	2,37	2,45	2,32
4	5066	2,20	2,20	2,26	2,20
5	1239	2,11	2,10	2,12	2,14
6	145	2,06	2,05	2,02	2,09
7	14	2,00	2,01	1,93	2,06
			R^2=.999	R^2=.977	R^2=.987

Abb. 4 veranschaulicht die Anpassungsgüte der beiden nicht-linearen Modelle.

Abb. 4: Nicht-lineare Anpassungsmodelle



Wir können somit festhalten, daß sich in dem untersuchen kroatischen Wörterbuchmaterial eine Korrelation zwischen Wort- und Silbenlänge feststellen läßt, die – wie die entsprechenden nicht-linearen Regressionsanalysen zeigen – mit zwei verschiedenen Regressionsgleichungen annähernd gleich gut zu erfassen ist. Beide Formeln lassen sich, wie Wimmer / Köhler / Altmann (Ms.) vorschlagen, aus einer gemeinsamen, übergeordneten Gleichung (5) ableiten, die als Grundlage weiterer Spezialfälle angesehen werden kann:

(5) $$y = Ce^{a_0/x} x^{a_1} e^{-a_2/x - a_3/(2x^2) - a_4/(3x^3) - \ldots}$$

Für $a_1 < 0, a_0 = a_2 = a_3 = \ldots = 0$ erhalten wir die oben dargestellte Formel (II) des Menzerath'schen Gesetzes:

(6) $$y = Cx^{a_1},$$

und die Gleichung

(7) $$y = e^{(a+bx)}$$

läßt sich nach dem Vorschlag von Wimmer / Köhler / Altmann (Ms.) für den Fall, daß $a_2 < 0, a_0 = a_1 = a_3 = \ldots = 0,$ umformulieren in:

(8) $$y = Ce^{-a_2/x}.$$

Formel (8) ist zwar u.a. zur Beschreibung des Zusammenhangs zwischen Wortlänge und Bedeutungsumfang angewandt worden, jedoch hat man Zusammenhänge von Konstrukt und Komponenten bislang eher mit der Formel (II) des Menzerath'schen Gesetzes (6) beschrieben. Insofern scheint es nicht uninteressant, in Zukunft weiter zu verfolgen, unter welchen Bedingungen beide Gleichungen effektiv sind und unter

welchen eine der beiden besser paßt (vgl. Grzybek 2000a). Der Frage, inwiefern die beobachteten Zusammenhänge auf ähnliche oder andere Art und Weise in Texten modelliert werden können, in denen zusätzliche die Wortlänge beeinflussende Faktoren wie Satzlänge und textspezifische Parameter ins Spiel kommt, muß an anderer Stelle nachgegangen werden (vgl. Grzybek 2000b, c).

Schließen wir mit diesen Bemerkungen unsere Überlegungen zur Wortlänge und Faktoren, mit denen sie korreliert, ab. Ohne Frage bleibt noch Einiges zu tun, bis wir die Gesetze verstehen, denen Korrelationen wie die oben aufgezeigten folgen. Doch die Kausalität dieser Korrelationen zu verstehen, würde der wissenschaftstheoretischen Notwendigkeit nach der Erklärung zugrundeliegender Mechanismen gleichkommen.

Literatur

Altmann, G. (1980): Prolegomena to Menzerath's Law. In: Grotjahn, R. (ed.), Glottometrika 2. – Bochum. (1-10).

Altmann, G.; M. H. Schwibbe (1989): Das Menzerathsche Gesetz in informationsverarbeitenden Systemen. – Hildesheim u.a.

Bunge, M. (1967): Scientific Research II: The Search for Truth. – New York.

Diehl, J. M.; H. U. Kohr (1979): Deskriptive Statistik. – Frankfurt a.M., ^{3}1979.

Gajić, D. M. (1950): Zur Struktur des serbokroatischen Wortschatzes. Die Typologie der serbokroatischen mehrsilbigen Wörter. Diss. – Bonn.

Grzybek, P. (2000a): Zur Worthäufigkeit, Wortlänge und Wortlängenhäufigkeit im Slowenischen. (Wortlistenanalysen). [Im Druck]

Grzybek, P. (2000b): Satzlängenverteilungen in slowenischen Sprichwörtern. [Im Druck]

Grzybek, P. (2000c): Wort- und Satzlänge im Kroatischen und Slowenischen. [Im Druck]

Sachs, L. (1992): Angewandte Statistik. Anwendung statistischer Methoden. – Heidelberg u.a, ^{7}1992.

Wimmer, G.; R. Köhler; G. Altmann (Ms.): Unified derivation of some linguistic laws. In: Handbook of Quantitative Linguistics.

Peter Grzybek (Graz)

Some marginal remarks on the correlation between word length and syllable length in Croatian

Based on the assumption that the notion of 'correlation' tends to be used in an increasingly blurred manner, an operational understanding of correlation is presented which is used in the field of statistics. Consequently, a correlation concerns the dependence between the extension of two features (variables) of a given individual, process, etc. By way of one example, the correlation between word length and syllable length in Croatian, it is shown that simple linear regression models are not adequate for its theoretical description. Rather, more complex non-linear regression models turn out to be efficient, two of which are compared to each other. Since only lexicon material is analyzed here, it is emphasized that different models are likely to be more adequate for the description of such linguistic material, which additionally includes sentence and/or text structures as influencing parameters.

Peter Grzybek
Institut für Slawistik der Universität Graz, Merangasse 70/I, A–8010 Graz
Tel: ++43 316 380-2526; Fax: ++43 316 380-9773
e-mail: grzybek@kfunigraz.ac.at