

Syntaktische Strukturen: Eigenschaften und Zusammenhänge

1 Einführung

In diesem Beitrag wird nach dem Muster der bereits aufgestellten und erfolgreich überprüften synergetisch-linguistischen Modelle im Bereich der Lexik (Köhler 1986; Hammerl 1991; Giesecking 2003) und der Morphologie (Köhler 1990a, 1990b, 1991; Krott 1996, 2003) ein Basismodell eines syntaktischen Subsystems der Sprache erstellt und empirisch überprüft. Für die theoretische Modellierung wird von zunächst einigen wenigen syntaktischen Einheiten, Eigenschaften und Zusammenhängen ausgegangen, die in ein entsprechendes Modell integriert werden. Die empirische Überprüfung erfolgt an Daten, die aus dem „Susanne-Korpus“ (Sampson 1995) für das Englische und dem „Negra-Korpus“ (Brants 1999) für das Deutsche gewonnen wurden.¹

Als Grundeinheit wählen wir die *syntaktische Konstruktion*, die hier auf der Basis der Konstituenz-Relation als *Konstituenten(typ)* operationalisiert wird. Die im Rahmen dieser Untersuchung betrachteten Eigenschaften sind:

- *Frequenz* (Aufretenshäufigkeit im Textkorpus, in Textsorten bzw. in Einzeltexten),
- *Länge* (Anzahl der terminalen Knoten bzw. Wörter, die zu einer Konstituente gehören),
- *Komplexität* (Anzahl der unmittelbaren Konstituenten der betrachteten Konstituente),
- *Position* (die von vorn nach hinten bzw. von links nach rechts inkrementierte laufende Nummer der betrachteten Einheit in Bezug auf die Mutterkonstituente bzw. auf den Satz),
- *Verschachtelungstiefe* (Anzahl der Ableitungsschritte vom Startsymbol bis zur betrachteten Konstituente),
- *Information* (im informationstheoretischen Sinn; entspricht u.a. dem Gedächtnisspeicherplatz, der zur Zwischenspeicherung der grammatischen Bezüge der betrachteten Konstituente benötigt wird),
- *Polyfunktionalität* (Anzahl der verschiedenen Funktionen, für die die betrachtete Konstruktion verwendet werden kann),
- *Synfunktionalität* (Anzahl der Funktionen, mit der eine gegebene Funktion ein syntaktisches Ausdrucksmittel teilt)

und die relevanten Inventare, nämlich

¹ Für ihre Hilfe bei der Gewinnung und Aufbereitung der Ausgangsdaten danke ich Claudia Prün und Sabine Weber.

- das Inventar der *syntaktischen Konstruktionen* (Konstituententypen),
- das Inventar der *syntaktischen Funktionen*,
- das Inventar der *syntaktischen Kategorien* (einschließlich der Wortarten),
- das Inventar der *funktionalen Äquivalente* (also der Konstruktionen, die annähernd gleiche Funktionen ausüben können wie die betrachtete).

2 Frequenz, Komplexität und Länge

Der erste Schritt auf dem Weg zu einem Modell im Rahmen des synergetischen Ansatzes besteht im Aufstellen von Axiomen. Außer dem allgemeinen zentralen Axiom der Selbstorganisation und Selbstregulation übernehmen wir aus den früheren Arbeiten (vgl. z.B. Köhler 1986, 1990a; Hoffmann/Krott 2003) das Kommunikationsbedürfnis (Kom) mit seinen beiden Aspekten: dem Kodierungsbedürfnis (Kod) und dem Anwendungsbedürfnis (Anw); weitere Sprach-externe Anforderungen an das System werden weiter unten eingeführt. Der nächste Schritt umfasst die Suche nach funktionalen Äquivalenten, die die Anforderungen bedienen können, und die Bestimmung ihrer Auswirkungen auf andere Größen des Systems.

Der Einfluß von Kod, von dem wir hier nur den Teil betrachten, der sich auf das funktionale Äquivalent der syntaktischen Ausdrucksmittel bezieht, betrifft direkt die Größe des Inventars syntaktischer Konstruktionen (ganz analog zum Einfluss von Kod im lexikalischen Subsystem, wo es die Lexikongröße betrifft). Anw, ebenfalls analog zur entsprechenden Wirkung in der Lexik, repräsentiert die kommunikative Relevanz eines inventarisierten Ausdrucks und resultiert in einer von dieser abhängigen Anwendungshäufigkeit dieser Konstruktion (vgl. Abb. 1).

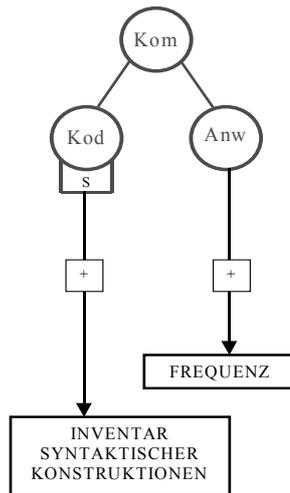


Abb. 1: Das sprachkonstituierende Bedürfnis Kod (nur die syntaktischen Ausdrucksmittel S werden betrachtet) und das sprachformende Bedürfnis Anw mit abhängigen Größen im syntaktischen Subsystem.

Vor dem nächsten Schritt, der empirischen Überprüfung der mit den vorangehenden Schritten aufgestellten Hypothesen, führen wir noch ein weiteres Axiom ein, nämlich das Bedürfnis nach *Optimierung der Kodierung* (OK) mit zwei seiner vielen Aspekte: dem (bereits in früheren Teilmodellen eingeführten) Bedürfnis nach *Minimierung des Produktionsaufwands* (minP) und dem nach *Maximierung der Kompaktheit* (maxK). Der Produktionsaufwand betrifft den physischen Aufwand, der mit der Artikulation einer Äußerung des gegebenen Ausdrucks verbunden ist. Im Falle syntaktischer Konstruktionen ist er durch die Anzahl der terminalen Knoten, der Wörter, bestimmbar (auch wenn die Wörter verschiedene Länge besitzen)² und soll *Länge* der syntaktischen Konstruktion heißen. MinP wirkt, wiederum ähnlich wie im Falle lexikalischer Einheiten, auf den Zusammenhang zwischen Frequenz und Länge; wie dort wird die maximale Einsparung an Aufwand erreicht, wenn die häufigsten Konstruktionen die kürzesten sind (vgl. Abb. 2). Entsprechend sind auch eine optimierte Häufig-

² Mit Hilfe der Wortlängenverteilung (in Silben) und der Verteilung der Silbenlänge (in Phonemen) ist der tatsächliche durchschnittliche Aufwand für die Äußerung einer syntaktischen Konstruktion indirekt durch die Zahl der Wörter gegeben. Zu beachten ist allerdings auch die Wirkung des Menzerath-Altmann-Gesetzes, was wir aber hier aus Gründen der Vereinfachung vernachlässigen.

keitsverteilung der Frequenzklassen und eine dazugehörige Rang-Frequenz-Verteilung zu erwarten, für die ein ähnlicher, wenn auch nicht identischer Verlauf wie bei dem Zipf-Mandelbrotschen Gesetz anzunehmen ist. Zwar gibt es wie bei Wörtern auch einen mit der Frequenz verbundenen Kürzungseffekt auf die Länge syntaktischer Konstruktionen; hauptsächlich aber ergibt sich der Zusammenhang aus der überwiegenden Verwendung kürzerer gegenüber längerer Konstruktionen.

Für beide Korpora findet sich tatsächlich die erwartete Form dieser Verteilungen (s. Abb. 3). An das Frequenzspektrum lässt sich die Waring-Verteilung anpassen (die Anpassung mit dem Altmann-Fitter 2.0 (1997) ergab z.B. für das Susanne-Koprus die geschätzten Parameter $b = 0,66889$ und $n = 0,47167$ bei 85 Freiheitsgraden. Es ergibt sich ein $\chi^2 = 81,0102$ mit einer Wahrscheinlichkeit $P[\chi^2] = 0,6024$).

Die Anforderung maxK ist u.a. eine Konsequenz aus dem Bedürfnis nach Minimierung des Produktionsaufwands auf der Ebene der Mutterkonstituenten. Die Bedienung dieser Anforderung auf der Satzebene kann z.B. dadurch erfolgen, dass anstelle eines Nebensatzes ein zusätzliches Attribut in eine der Nominalphrasen eingefügt wird.³ Die Länge (gemessen in Wörtern) wiederum ist stochastisch proportional zur Komplexität: Je mehr unmittelbare Konstituenten eine Konstruktion besitzt, desto mehr terminale Knoten muss sie schließlich haben.

³ Beispiel: s[NP[Die SeminarteilnehmerInnen] konnten nichts verstehen, weil sie wieder einmal nicht vorbereitet waren] → s[NP[Die wieder einmal nicht vorbereiteten SeminarteilnehmerInnen] konnten nichts verstehen]. Der erste Satz ist 12 Wörter lang, der zweite nur 9, dafür hat das Subjekt des ersten Satzes nur zwei unmittelbare Konstituenten und eine Länge von 2 Wörtern, das des zweiten drei unmittelbare Konstituenten und eine Länge von 6.

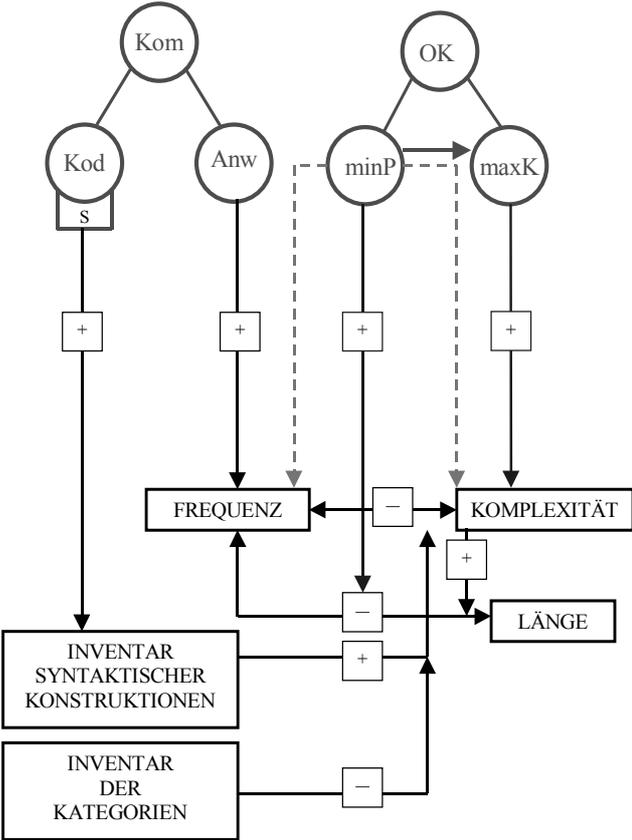


Abb. 2: Der Zusammenhang von Komplexität und Länge mit der Frequenz aufgrund der Anforderungen nach Optimierung der Kodierung. Die gestrichelten Linien repräsentieren die Wirkung von minP als Ordnungsparameter für die Verteilung der Häufigkeiten und der Komplexitätsklassen.

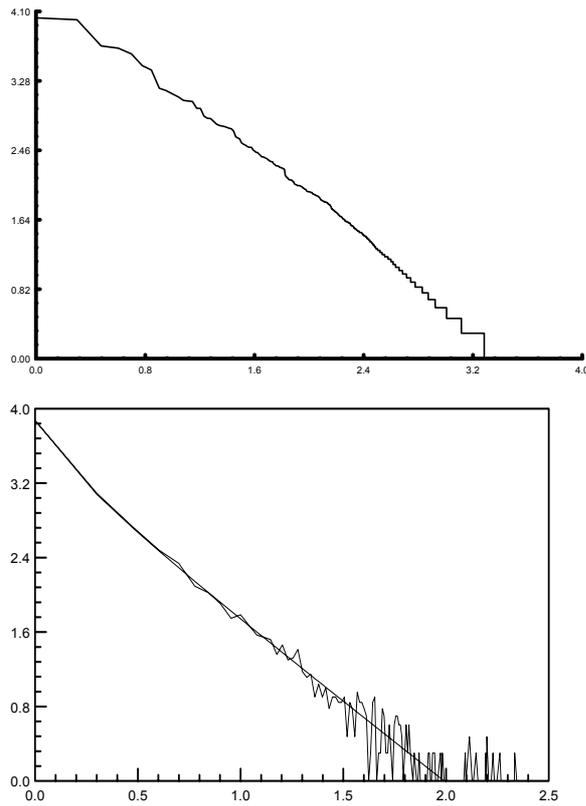


Abb. 3: Die Rang-Frequenz-Verteilung der Konstituentenhäufigkeiten im Susanne-Korpus (links); Spektraldarstellung der Konstituentenhäufigkeiten im Negra-Korpus und Anpassung der Waring-Verteilung (rechts). Beides in doppelt logarithmischer Darstellung.

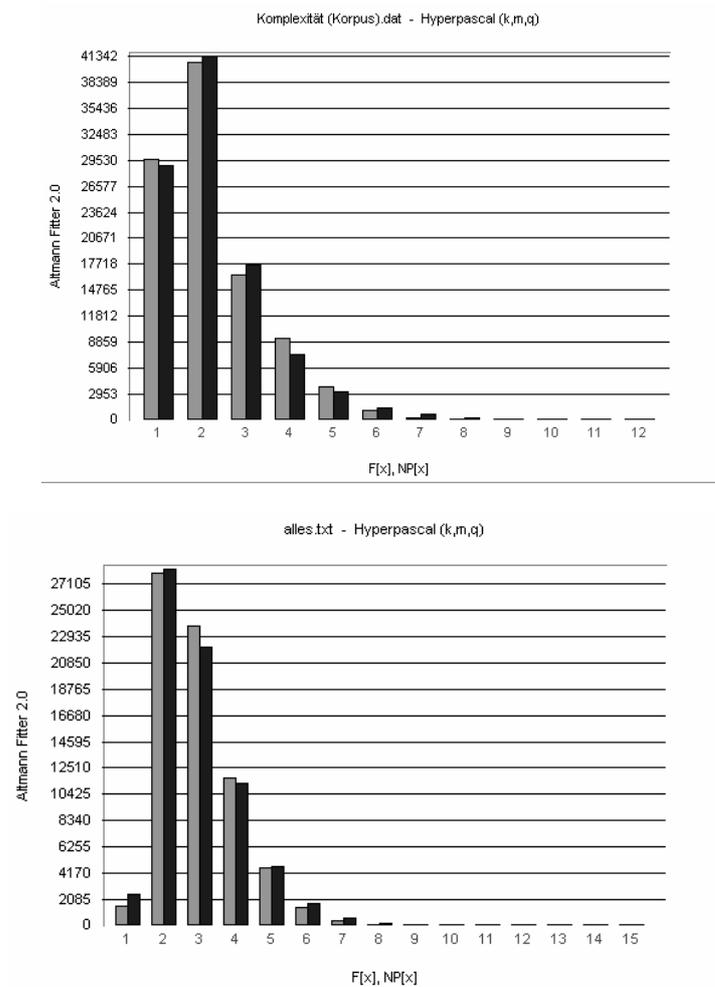


Abb. 4: Die empirische Häufigkeitsverteilung der Konstituentenkomplexität im Susanne-Korpus (links) und im Negra-Korpus (rechts), jeweils mit Anpassung der Hyperpascal-Verteilung.

Die *durchschnittliche* Komplexität der syntaktischen Konstruktionen schließlich ist von der Zahl der benötigten und inventarisierten Konstruktionen und der Anzahl der elementaren syntaktischen Kategorien abhängig. Dieser Zusammenhang ergibt sich aus einfacher Kombinatorik: Jede Konstruktion besteht aus einer linearen Folge von

Tochterknoten (unmittelbaren Konstituenten) und ist durch deren Kategorien und Reihenfolge bestimmt. Mit einer gegebenen Anzahl G von Kategorien lassen sich G^K verschiedene Konstruktionen mit K Knoten erzeugen, von denen allerdings nur ein Teil (der „grammatische“) tatsächlich gebildet wird – analog zur nicht vollständigen Ausnutzung der prinzipiell möglichen Phon(em)kombinationen bei der Bildung von Silben (Morphen), die zur Phonotaktik führt. Abbildung 4 zeigt die Verteilung der Komplexität sämtlicher 101138 Konstituentenvorkommen im Susanne-Korpus und die der 71494 Vorkommen im Negra-Korpus.

Die empirische Überprüfung der Hypothesen über den Zusammenhang zwischen Frequenz und Komplexität und Komplexität und Länge ist in den Abbildungen 5, 6 und 7 dargestellt.

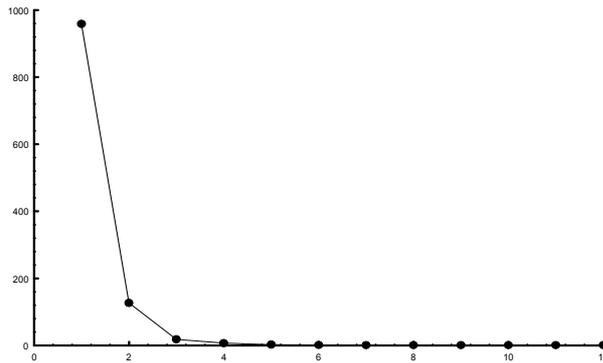


Abb. 5: Die empirische Abhängigkeit der mittleren Frequenz der Konstituenten als Funktion ihrer Komplexität. Anpassung der Funktion $F = 858,83 K^{-3,095} e^{0,00727K}$ mit dem Determinationskoeffizienten $D = 0,99$.

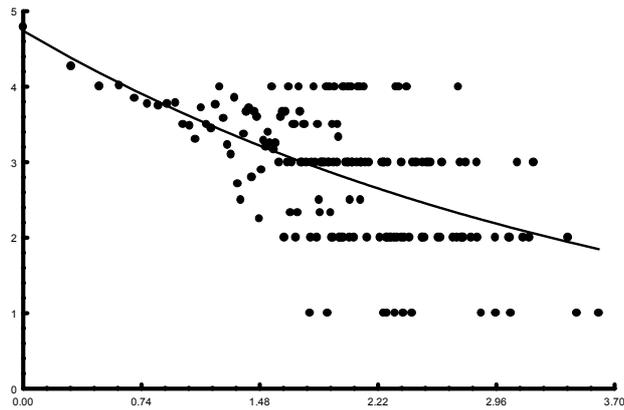


Abb. 6: Die empirische Abhängigkeit der mittleren Komplexität als Funktion der Frequenz und Anpassung der Funktion $K = 4,789 F^{-0,1160}$ (Determinationskoeffizient $D = 0,331$). Die x-Achse wurde logarithmiert.

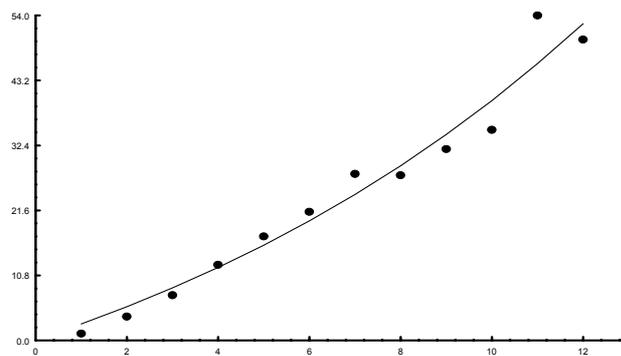


Abb. 7: Der empirische Zusammenhang zwischen Komplexität und Länge. Anpassung der aus der Modellstruktur abgeleiteten Funktion $L = 2,603 K^{0,963} e^{0,0512 K}$ mit dem Determinationskoeffizienten $D = 0,960$.

Die vermuteten Tendenzen bestätigen sich offensichtlich. Obwohl (außer im Fall der Abhängigkeit der Länge von der Komplexität) noch keine theoretisch begründeten Hypothesen über die mathematische Form der Zusammenhänge vorliegen und daher kein ernsthafter Signifikanztest durchgeführt werden kann, muss man die generelle Hypothese über die Existenz einer inversen Abhängigkeit aufgrund der Daten als empirisch haltbar ansehen.

Diesen Befunden, besonders dem Frequenzspektrum, kommt potenziell auch eine wichtige praktische Bedeutung zu: Von den 4621 verschiedenen Konstituententypen mit 90821 Vorkommen im gesamten Susanne-Korpus kommen 2710 Typen (58,6%) nur ein einziges Mal vor, von den restlichen 1911 Typen kommen 615 (32,3% = 13,3% vom Gesamtinventar) zweimal, von den dann verbleibenden 1296 Typen 288 (22,2% = 6,2 % vom Gesamtinventar) dreimal, von den danach 1008 übrigen Typen 176 viermal (17,5% = 3,8% vom Gesamtinventar) vor etc. Weniger als 20% der entsprechenden Grammatikregeln sind also öfter als viermal, weniger als 30% der Regeln öfter als zweimal anwendbar (vgl. Tabelle 1).

Frequenz	Anzahl (Rest)	% vom Gesamtinventar	% vom jeweiligen Rest
1	2710 (4621)	58,6	—
2	615 (1911)	32,3	13,3
3	288 (1296)	22,2	6,2
4	176 (1008)	17,5	3,8

Tabelle 1: Die Belegung der Frequenzklassen von Konstituententypen im Susanne-Korpus.

Es ist zu erwarten, dass die Untersuchung anderer Korpora, auch anderer Sprachen als des Englischen und des Deutschen, grundsätzlich vergleichbare Ereignisse bringen wird. Ähnlich wie die Frequenzspektra des Vokabulars seit langem in der Sprachlehrforschung, bei der Erstellung von Grund- und Minimalwortschätzen, bei der Strukturierung von Sprachlehrmaterial und der Zusammenstellung von Wörterbüchern (bzw. der Bestimmung ihres Textabdeckungsgrads) Verwendung finden, könnte die oben dargestellte Gesetzmäßigkeit u.a. beim Schreiben von Grammatiken und Konstruieren von Parsern, Planung des Abdeckungsgrads, Abschätzung des Aufwands für die Regelerstellung, Berechnung der automatischen Analysierbarkeit von Texten z.B. in der linguistischen Datenverarbeitung usw.) berücksichtigt werden.

3 Länge, Komplexität und Position

Ein bereits von Otto Behaghel in (Behaghel 1930) festgestellter Zusammenhang ist das von ihm so genannte *Gesetz der wachsenden Glieder*: „Von zwei Gliedern von verschiedenem Umfang steht das umfangreichere nach“, das er an Sprachdaten des Deutschen, Lateinischen und Griechischen nachprüfte. Wortstellungsvariation ist seitdem vor allem unter typologischen Gesichtspunkten betrachtet worden. In der Linguistik sind besonders die Thema-Rhema-Gliederung und Topikalisierung als Funktion syntaktischer Kodierung mittels Wortstellung zu nennen sowie demgegenüber Givóns diskurspragmatisches Prinzip „das Wichtigste zuerst“. Eine interessante und plausible Hypothese zur Begründung der von Behaghel beobachteten und von ihm selbst am Deutschen, Englischen und Ungarischen überprüften Präferenz „lang hinter kurz“ von John Hawkins (1994) beruht auf psycholinguistischen Annahmen über die Mechanismen und Randbedingungen der menschlichen Sprachverarbeitung. Sein EIC-Prinzip („Early Immediate Constituent Principle“) erklärt die empirischen Befunde, dass grammatisch gleichberechtigte Konstituenten vorzugsweise so angeordnet werden, dass die längeren hinter den kürzeren stehen, damit, dass auf diese Weise bei der syntaktischen Analyse weniger Knoten im Strukturbaum zwischengespeichert werden müssen. Hoffmann (1996), die verschiedene empirische Tests durchführt, da die von Hawkins selbst erhobenen Daten methodisch nur unzureichend überprüft werden (so führt er keinen statistischen Signifikanztest durch, sondern bewertet die Daten intuitiv) zeigt darüber hinaus, dass die Wahrscheinlichkeit, mit der eine längere Konstituente hinter der kürzeren steht, eine monotone Funktion des Längenunterschieds ist: je größer der Unterschied, desto wahrscheinlicher ist eben diese Stellungsvariante. Neuere Untersuchungen zum Verhältnis Länge und Position von Wörtern findet man in (Uhlířová 1997a, 1997b). In Abbildung 8 wird dieser Zusammenhang grafisch in einer Form dargestellt, die seine Integration in ein umfassendes syntaktisches Teilmodell ermöglicht. Anstelle der Länge in Wortanzahl ist hier die Komplexität als relevante Größe angesetzt, da es ja nicht um zwischengespeicherte Wörter, sondern um Knoten im Syntaxbaum geht. Dass sich die Wirkung des Prinzips auch an der Länge in Wörtern nachweisen lässt, ist offensichtlich ein indirekter Effekt.

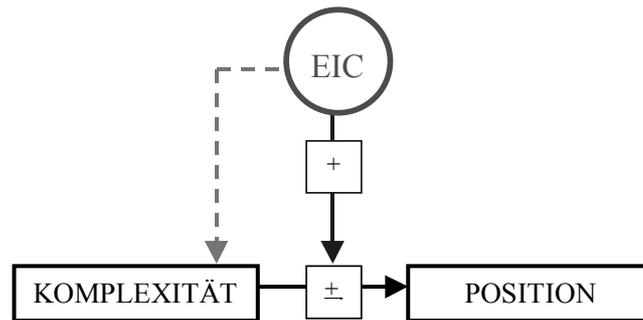


Abb. 8: Hawkins' EIC-Prinzip modifiziert (Komplexität anstelle von Länge).

Zur nochmaligen Überprüfung dieser Hypothese wurden ebenfalls Daten aus dem Susanne-Korpus herangezogen. Abweichend von den anderen Untersuchungen wurde nicht die Stellung ‚gleichwertiger Konstituenten‘ verschiedener Länge paarweise verglichen, sondern es wurden die Daten zu Länge, Komplexität und absoluter Position in der jeweiligen Mutterkonstituente sowohl für alle Satzkonstituenten als auch für sämtliche Konstituenten auf allen Verschachtelungsebenen erhoben und ausgewertet. Beispiele für die empirisch gefundenen Zusammenhänge sind in den Abbildungen 9 und 10 dargestellt. Wie man sieht, ergab sich eine klare Bestätigung der Hypothese; die Abhängigkeiten sind so deutlich, dass sich ein Signifikanztest erübrigt, zumal zur Zeit noch keine theoretische Begründung für die Annahme einer speziellen Funktion vorliegt, die mittels Anpassung an die Daten geprüft werden könnte. Bei der Entwicklung einer entsprechenden Formel werden außer Hawkins' Überlegungen auch andere Zusammenhänge wie Givóns Prinzip „das Wichtigste zuerst“ und die quantitative Ikonizität (z.B. Haiman: „je wichtiger, desto mehr sprachliches Material“) zu berücksichtigen und miteinander zu verrechnen sein.

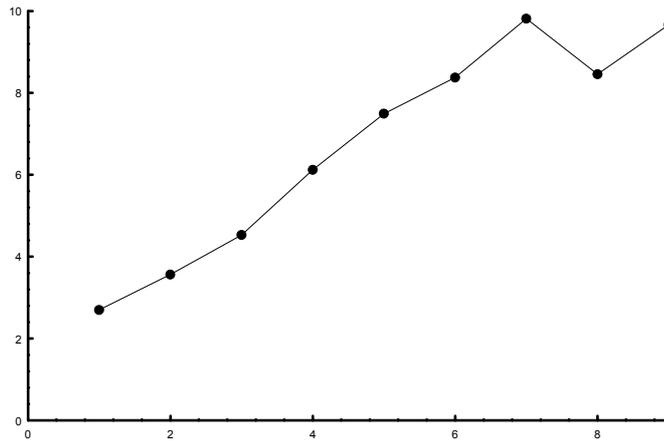


Abb. 9: Die empirische Abhängigkeit der mittleren Konstituentenlänge (in Wörtern) und der Position der jeweiligen Konstituente in der Mutterkonstituente. Die Werte für Positionen oberhalb von 8 wurden wegen mangelnder Belegzahl (<10) nicht berücksichtigt.

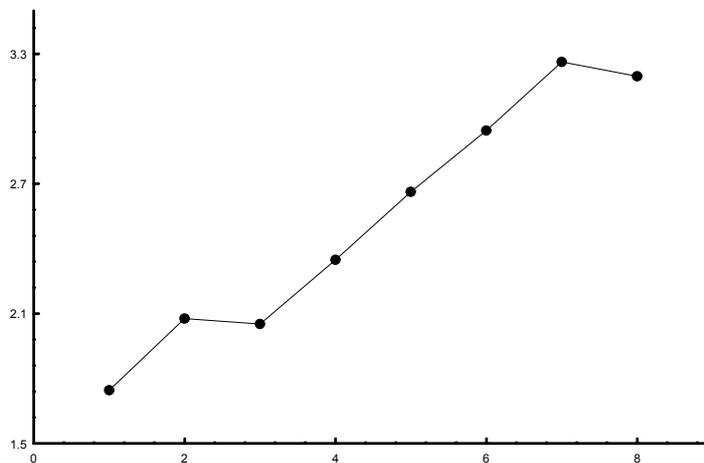


Abb. 10: Die empirische Abhängigkeit der mittleren Konstituentenkomplexität (in unmittelbaren Konstituenten) und der Position der jeweiligen Konstituente in der Mutterkonstituente. Die Werte für Positionen oberhalb von 8 wurden wegen mangelnder Belegzahl (<10) nicht berücksichtigt.

4 Position und Verschachtelungstiefe

Als weitere Hypothese, zu der Daten leicht zu erfassen sind, integrieren wir eine Konsequenz aus Yngves (1960) „Depth Saving Principle“. Wenn zutrifft, dass aus Gründen der Verarbeitungseffizienz Rechtsverzweigungen vorgezogen werden, sollte sich bei allen Konstituenten mit wachsender Position eine Zunahme der durchschnittlichen Verschachtelungstiefe⁴ zeigen. Um dies zu überprüfen, wurden Verschachtelungstiefe (die Satzebene wurde als Tiefe 1 gezählt) und absolute Position (in der Mutterkonstituente und, separat, im Satz) für alle Konstituentenvorkommen im Korpus ausgewertet. Der empirische Zusammenhang dieser beiden Variablen ist in Abbildung 11 dargestellt. Abbildung 12 zeigt den Zusammenhang zwischen Verschachtelungstiefe und Position im Satz.

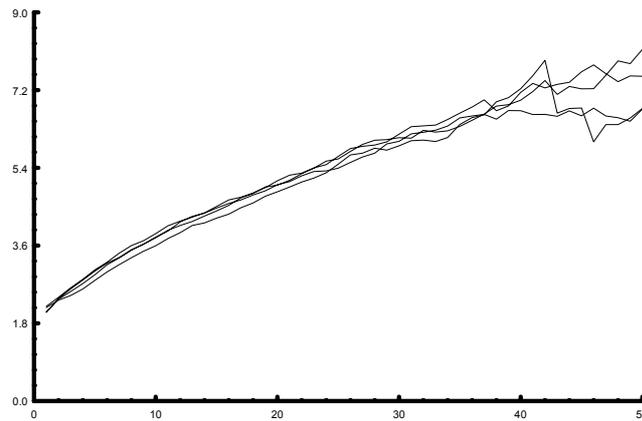


Abb. 11: Die empirische Abhängigkeit der Verschachtelungstiefe von Konstituenten von ihrer Position (gemessen in laufenden Wörtern vom Satzbeginn) für vier Teilkorpora (Texttypen) gesondert. Daten oberhalb der Wortposition 50 sind wegen geringer Belegungsdichte unzuverlässig und wurden nicht dargestellt.

⁴ Es sei darauf hingewiesen, dass sich die hier verwendete Operationalisierung des Begriffs Verschachtelungstiefe von der in Yngve (1960) unterscheidet.

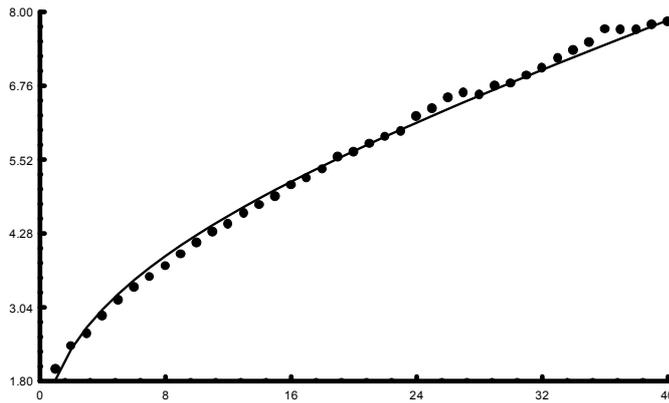


Abb. 12: Die empirische Abhängigkeit der Verschachtelungstiefe von Konstituenten von ihrer Position (gemessen in laufenden Wörtern vom Satzbeginn) für das Gesamtkorpus. Daten oberhalb der Wortposition 40 sind wegen geringer Belegungsdichte unzuverlässig und wurden nicht dargestellt. Anpassung der Funktion $T = 1,8188 P^{3,51} e^{0,00432 P}$ mit dem Determinationskoeffizienten $D = 0,996$.

5 Position und Information

In Bezug auf die Position von Konstituenten lässt sich eine weitere Überlegung anstellen: Sie ist ein Maß für die Zahl der bei der Verarbeitung durch den Hörer/Leser zwischenspeichernden Konstituenten, also ein Maß auch für die Auslastung des Gedächtnisspeichers. Nimmt man an, dass außer den Knoten selbst auch die Strukturinformation (vgl. Köhler 1984), die als Resultat der Analyse sukzessiv aufgebaut wird, in dem Verarbeitungsregister gehalten werden muss, dann wäre es von Vorteil, wenn die Menge an zu speichernder Strukturinformation nicht in demselben Ausmaß zunimmt wie die Knotenzahl, um möglichst komplexe Konstituenten verarbeiten zu können. Erreichbar wäre das, wenn die Zahl der jeweils grammatisch zulässigen alternativen Konstituententypen bzw. -funktionen mit wachsender Position sinkt. Je mehr Alternativen zulässig sind, desto mehr Speicherplatz ist ja zur Kodierung des tatsächlich gefundenen Typs erforderlich. Wenn die hier angestellte Überlegung zutrifft, wenn also die Sprachen mit Hilfe ihrer Selbstorganisationsmechanismen ihre Grammatiken so einrichten – bzw. das Sprachverhalten dafür sorgt –, dass eine Maximie-

rung der Komplexität der verarbeitbaren Konstituenten auf diese Weise erreicht wird, sollte der Logarithmus (als Maß für die Größe des erforderlichen Codes) der Zahl der in einem Textkorpus zu findenden Alternativen mit der Position sinken. In Abbildung 14 ist diese Hypothese in das Modell aufgenommen worden; das entsprechende Bedürfnis nach Minimierung der Strukturinformation (minS) ist ebenfalls ein Aspekt der allgemeineren Anforderung minG.

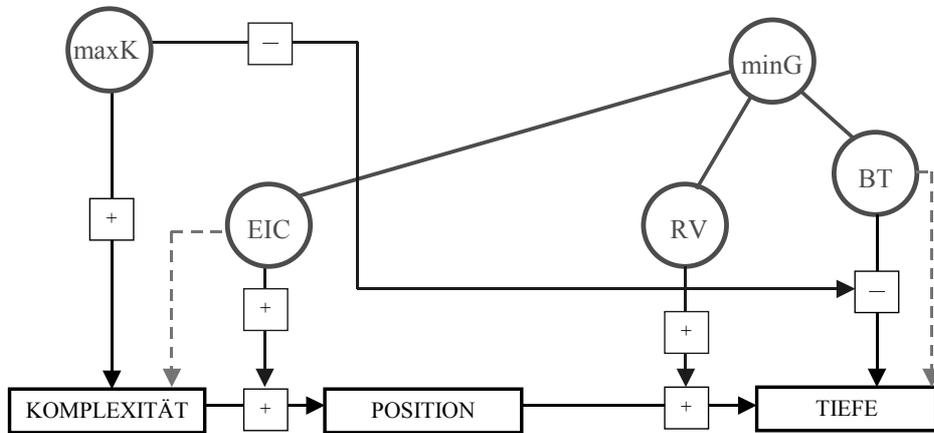


Abb. 13: Modellausschnitt mit den Größen Komplexität, Position und Tiefe und den dazugehörigen Anforderungen.

Für die empirische Prüfung wurden wieder zwei Datensätze untersucht, von denen einer alle Konstituenten auf Satzebene, der andere sämtliche Konstituenten auf allen Ebenen enthielt. Für alle Positionen in Bezug auf die Mutterkonstituente wurde ermittelt, wie viele verschiedene Nachfolgekonstituenten jeweils vorkamen. Auf der Satzebene ergibt sich hinter der Position 2 (der bevorzugten Position des finiten Verbs, das von der für die Analyse im Susanne-Korpus verwendeten Grammatik als unmittelbare Satzkonstituente angesehen wird) tatsächlich ein nahezu lineares Absinken der Strukturinformation (vgl. Abb. 15). Das Susanne-Korpus enthält überall, wo es möglich war, neben den Annotaten zu den Konstituententypen auch Information über die grammatische Funktion der betreffenden Konstituenten (vgl. Sampson 1995), so dass es möglich war, auch die Zahl der alternativen Funktionen je Position zu erheben. Auch dabei zeigt sich (vgl. Abb. 16a und 16b) ein fast linearer Verlauf des Informationswerts.

Zurzeit liegen keine Hypothesen über die exakte mathematische Form dieser funktionalen Abhängigkeiten vor. Solche können sinnvoll nur aus theoretischen Annahmen über den menschlichen Sprachverarbeitungsapparat und unter Berücksichtigung von Gesetzen wie dem Menzerath-Altmann-Gesetz gewonnen werden (s. Köhler 1984, wo eben dieses aus einfachen Überlegungen zu Eigenschaften der Sprachverarbeitung abgeleitet wird).

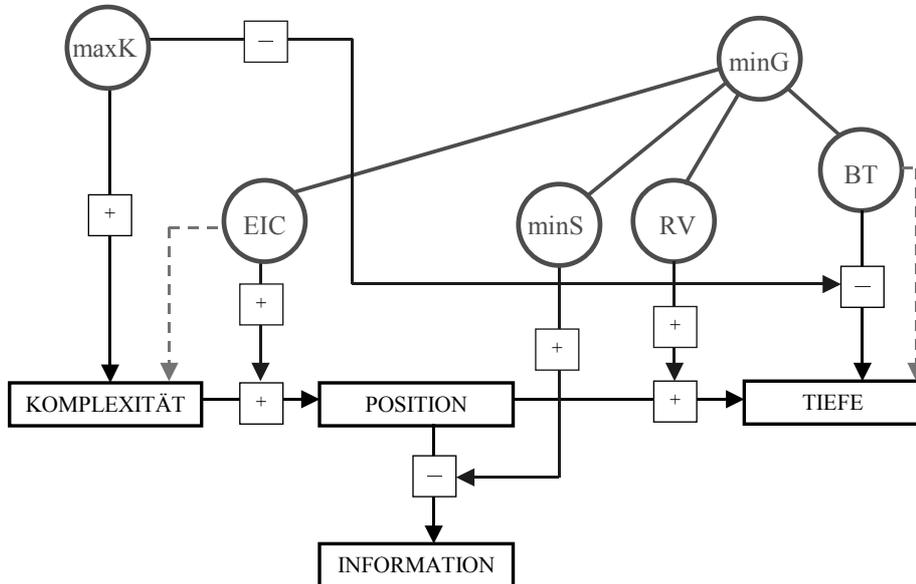


Abb. 14: Mit wachsender Position sinkt die Menge an zu speichernder Strukturinformation.

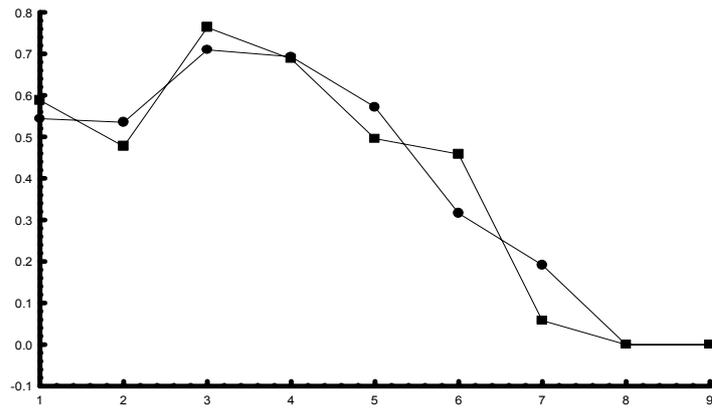


Abb. 15: Logarithmus der Anzahl alternativ möglicher Konstituententypen in Abhängigkeit von der Position (für zwei der vier Textsorten im Susanne-Korpus gesondert berechnet).

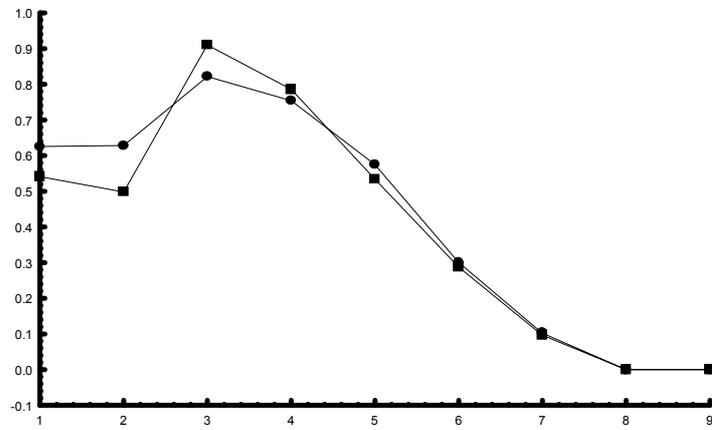


Abb. 16a: Logarithmus der Anzahl alternativ möglicher Konstituentenfunktionen in Abhängigkeit von der Position (für zwei der vier Textsorten im Susanne-Korpus gesondert berechnet).

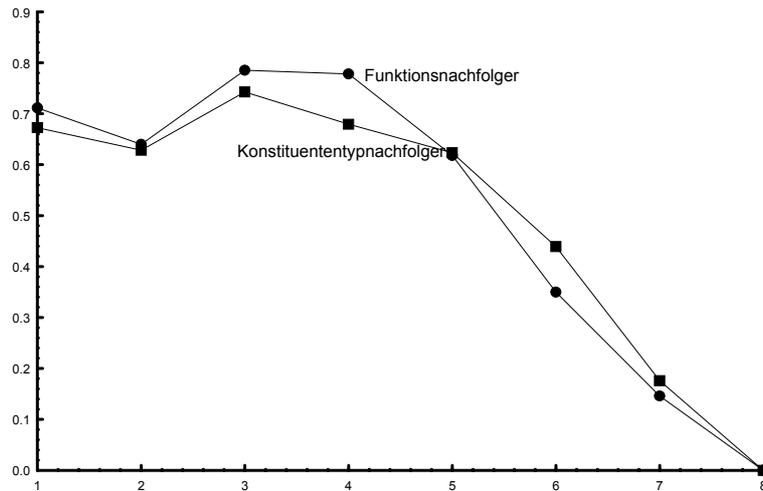


Abb. 16b: Logarithmus der Anzahl alternativ möglicher Konstituententypen bzw. Konstituentenfunktionen in Abhängigkeit von der Position (für einen einzelnen Text des Susanne-Korpus gesondert berechnet).

6 Weitere Größen und Anforderungen

Wie in anderen, früher aufgestellten Teilmodellen wird auch für das syntaktische Subsystem ein Systembedürfnis nach Inventarminimierung (minI) postuliert. Die Zusammenhänge zwischen den Inventargrößen sind folgende:

Wie schon gesagt, wirkt eine Vergrößerung des Inventars der syntaktischen Konstruktionen erhöhend auf die durchschnittliche Komplexität der Konstruktionen, während die Größe des Kategorien-Inventars sie vermindert. Je geringer die Größe des Kategorien-Inventars, desto größer die durchschnittliche funktionale Belastung (oder Multifunktionalität). Das Bedürfnis minI wirkt sich auf alle Inventare verringernd aus, darunter auf die Zahl funktionaler Äquivalente, die einer Konstruktion durchschnittlich zukommt. Die Häufigkeitsverteilungen innerhalb der Inventare werden durch Ordnungsparameter gesteuert (vgl. Abb. 17).

Die theoretische und empirische Analyse der Wahrscheinlichkeitsverteilungen der betrachteten Größen (Frequenz, Länge, Komplexität, Position, Tiefe und Infor-

mation) und der Rang-Frequenz-Verteilungen der Inventare wurde in (Köhler/Altmann 2000) durchgeführt. Hier soll lediglich angemerkt werden, dass die empirischen Häufigkeitsverteilungen von funktionalen Äquivalenten und der Multifunktionalität einer gegebenen Konstruktion wie auch die Häufigkeitsverteilungen, die die Diversifikation von Funktionen in Bezug auf verschiedene Konstruktionen (Synfunktionalität) wiedergeben, sehr heterogen sind und eine große Zahl verschiedener Modelle (Wahrscheinlichkeitsfunktionen) nötig ist, um sie zu erfassen.

Das im vorliegenden Beitrag entwickelte Modell für den betrachteten Ausschnitt der Syntax ist, wie einleitend betont, lediglich ein erster Versuch, dieses Subsystem im Rahmen des synergetischen Ansatzes zu modellieren, der zurzeit in verschiedener Hinsicht noch unvollständig bleiben musste. Vor allem fehlt noch die theoretische Ableitung der mathematischen Form einiger der funktionalen Zusammenhänge. Außer der Erweiterung des Modells um weitere Einheiten und Eigenschaften wird vor allem die Verbreiterung der empirischen Basis um möglichst viele andere Sprachen zu leisten sein. Darüber hinaus wird der im Diagramm angedeutete Zusammenhang zwischen der dargestellten Modellstruktur und dem Menzerathschen (auch Menzerath-Altmann-) Gesetz auszuarbeiten sein.

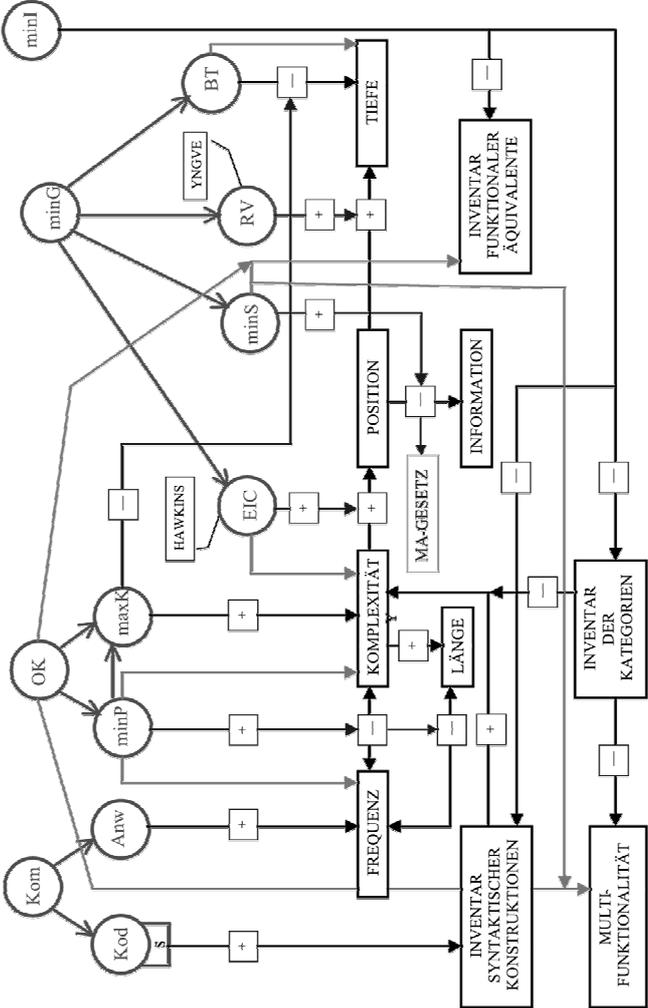


Abb. 17: Gesamtstruktur des betrachteten syntaktischen Subsystems.

Literatur

- Behaghel, Otto (1930): Von deutscher Wortstellung. *Zeitschrift für Deutschkunde*, 44, 81-89.
- Brants, Thorsten (1999): Tagging and Parsing with Cascaded Markov Models. Automation of Corpus Annotation. Saarbrücken: Universität des Saarlands.
- Gieseking, Kathrin. (2003): Untersuchungen zur Synergetik der englischen Lexik. In: Köhler, Reinhard (2003).
- Hammerl, Rolf (1991): Untersuchungen zur Struktur der Lexik: Aufbau eines Basismodells. Trier: Wissenschaftlicher Verlag Trier.
- Hawkins, John A. (1994): A performance theory of order and constituency. Cambridge: University Press.
- Hoffmann, Christiane (1996): Quantitativ-funktionalanalytische Untersuchungen zur Wortstellungsvariation. Magisterarbeit Trier.
- (1999): Word order and the principle of „Early Immediate Constituents (EIC)“. *Journal of Quantitative Linguistics*, 6, 101-107.
- /Krott, Andrea (2003): Einführung in die synergetische Linguistik. In: Köhler, Reinhard (2003).
- Köhler, Reinhard (1984): Zur Interpretation des Menzerathschen Gesetzes. In: Boy, Joachim/Köhler, Reinhard (Hrg.): Glottometrika 6. Trier: Wissenschaftlicher Verlag Trier. 177-183.
- (1986): Zur linguistischen Synergetik. Struktur und Dynamik der Lexik. Bochum: Brockmeyer (= Quantitative Linguistics 31).
- (1990a): Elemente der synergetischen Linguistik. In: Hammerl, Rolf (Hrg.): Glottometrika 12. Bochum: Brockmeyer, 179-187.
- (1990b): Synergetik und sprachliche Dynamik. In Koch, Walter A. (Hrg.): Natürlichkeit der Sprache und Kultur. Bochum: Brockmeyer, 96-112.
- (1991): Diversification of coding methods in grammar. In Rothe, Ursula (Hrg.): Diversification processes in language. Hagen: Rottmann Medienverlag, 47-55.
- (Hrg.) (2003): Beiträge zur quantitativen Korpuslinguistik (erscheint).
- /Altmann, Gabriel (2000): Probability Distributions of Syntactic Units and Properties. *Journal of Quantitative Linguistics* 7, 189-200.
- Krott, Andrea (1996). Some remarks on the relation between word length and morpheme length. *Journal of Quantitative Linguistics* 3, 29-37.
- (2003). Ein funktionalanalytisches Modell der Wortbildung. In: Köhler, Reinhard (2003).
- Sampson, Geoffrey (1995). English for the Computer. Oxford.
- Uhliřová, Ludmila (1997a): Length vs. Order. Word Length and Clause Length from the Perspective of Word Order. *Journal of Quantitative Linguistics* 4, 266-275.
- (1997b). O vztahu mezi délkou slova a jeho polohou ve větě. *Slovo a slovesnost*, 58, 174-184.
- Yngve, Victor (1960): A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society* 104, 444-466.