Reinhard Köhler (Hg.)
Korpuslinguistische Untersuchungen
zur quantitativen und systemtheoretischen Linguistik
2002, S. 365 - 385

Syntaktische Strukturen, Eigenschaften und Zusammenhänge

Reinhard Köhler

Einführung

In diesem Beitrag soll ein erster Versuch beschrieben werden, nach dem Muster der bereits aufgestellten und erfolgreich überprüften synergetisch-linguisti-schen Modelle im Bereich der Lexik (Köhler, 1986; Hammerl, 1991; Giese-king, 2002) und der Morphologie (Köhler, 1990a, 1990b, 1991; Krott, 1996, 2002) ein Basismodell eines im Rahmen der synergetischen Linguistik aufge-stellten syntaktischen Subsystems der Sprache zu erstellen und zu überprüfen. Für die theoretische Modellierung wird von zunächst einigen wenigen syntak-tischen 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) gewonnen wurden[1].

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

Frequenz (Auftretenshäufigkeit im Textkorpus),
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 inkriminierte laufende Nummer der betrachteten Einheit in bezug auf die Mutterkonstituente bzw. auf den Satz),

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

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 betrach-tete Konstruktion verwendet werden kann),
Synfunktionalität (Anzahl der Funktionen, mit der eine gegebene Funktion ein syntaktisches Ausdrucksmittel teilt)

und die relevanten Inventare, nämlich

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).

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 zen-tralen Axiom der Selbstorganisation und Selbstregulation übernehmen wir aus den früheren Arbeiten (vgl. z.B. Köhler, 1986, 1990a; Hoffmann & Krott, 2002) 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).

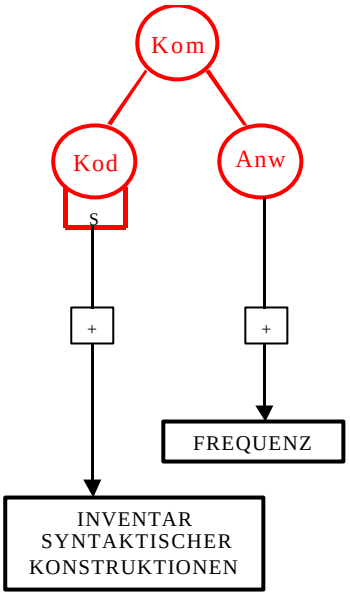


Abb. 1: Das sprachkonstituierende Bedürfnis **Kod** (nur die syntaktischen Ausdrucksmittel **S** werden betrachtet) und das sprachformende Bedürfnis Anw mit abhängigen Größen im syntaktischen Subsystem.

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 zweien seiner vielen Aspekte: dem (bereits in früheren Teilmodellen eingeführten) Bedürfnis nach *Minimierung des Produktionsaufwands* (minP) und dem nach *Maximierung der Kompaktheit* (maxK). Der Produktions-aufwand 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 syn-taktischen 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 Häufigkeitsverteilung der Frequenzklassen und eine dazugehörige Rang-Frequenz-Verteilung zu erwarten, für die ein ähnlicher, wenn auch nicht

[2] Mit Hilfe der Wortlängenverteilung (in Silben) und der Verteilung der Silbenlänge (in Phonen) 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.

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 Ver-wendung kürzerer gegenüber längerer Konstruktionen.

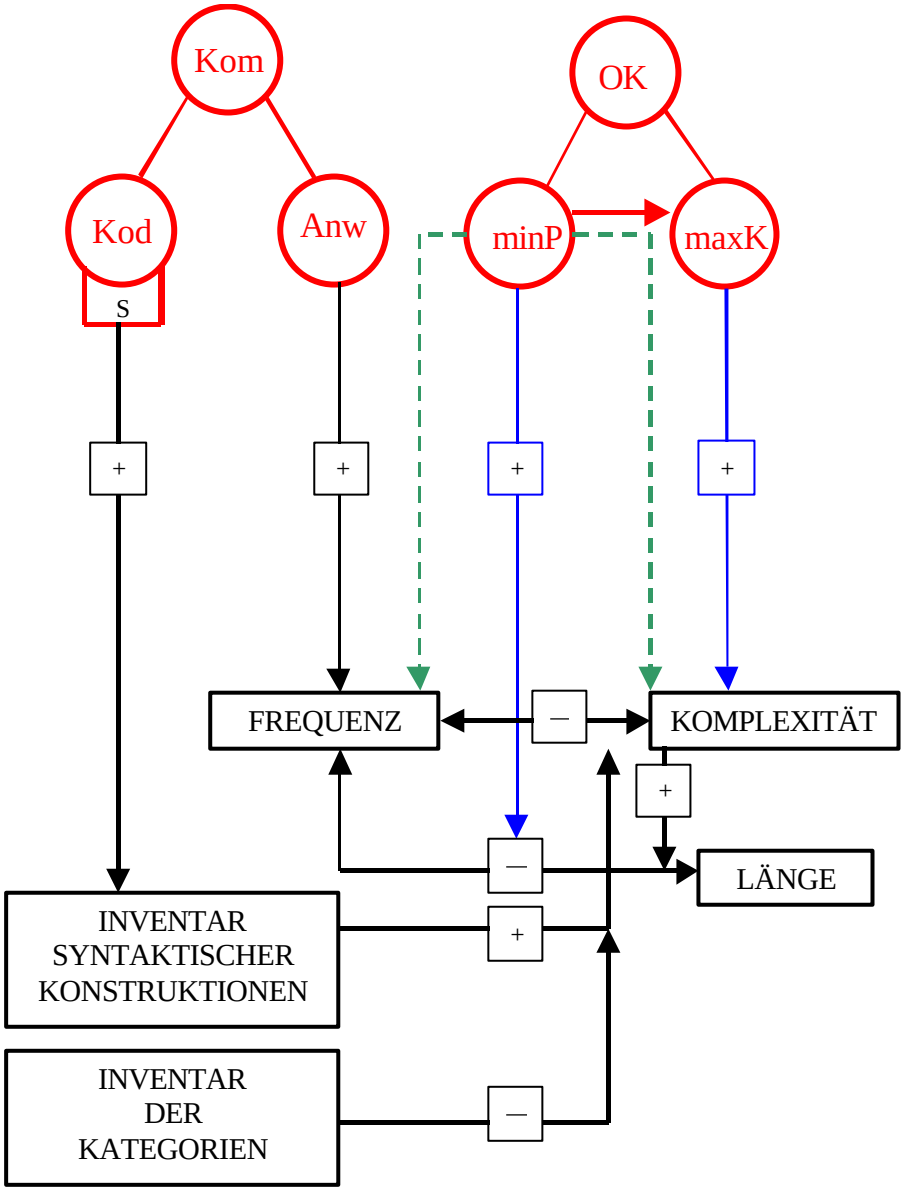


Abb. 2: Der Zusammenhang von Komplexität und Länge mit der Frequenz aufgrund der Anforderungen nach Optimierung der Kodierung. Die gestrichelten Linien repräsentieren die Wirkung von minP als Ord-nungsparameter für die Verteilung der Häufigkeiten und der Kom-plexitätsklassen.

Für das Susanne-Korpus findet sich tatsächlich die erwartete Form dieser Verteilungen (s. Abb. 3). An das Frequenzspektrum läßt sich die Waring-Ver-teilung anpassen (die Anpassung mit dem Altmann-Fitter 2.0 (1997) ergab die geschätzten Parameter b = 0,66889 und n = 0,47167 bei 85 Freiheitsgraden. Es ergibt sich ein X^2 = 81,0102 mit einer Wahrscheinlichkeit P[X^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 Konsti-tuenten eine Konstruktion besitzt, desto mehr terminale Knoten muß sie schließlich haben.

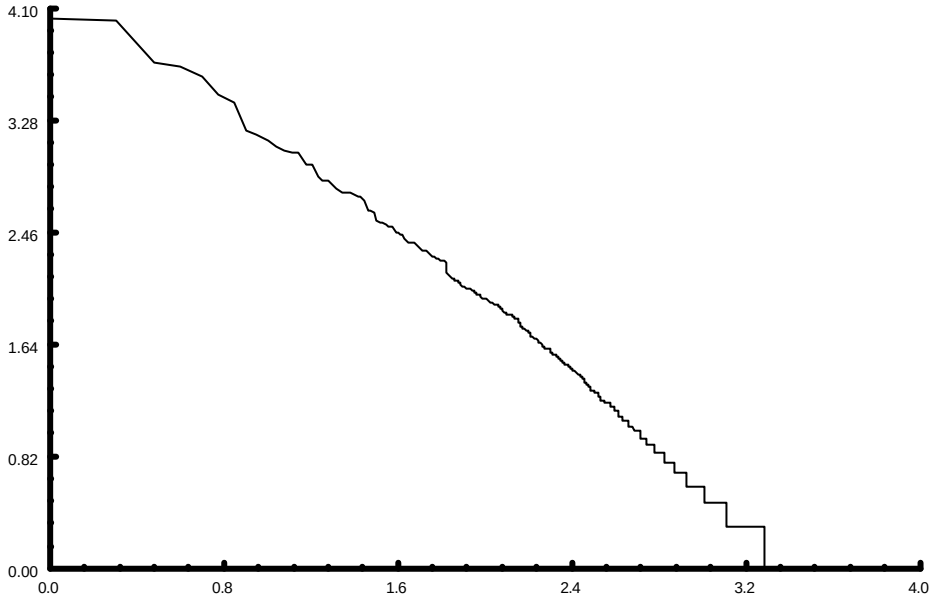


Abb. 3: Die Rang-Frequenz-Verteilung der Konstituentenhäufigkeiten im Susanne-Korpus in doppelt logarithmischer Darstellung

Die *durchschnittliche* Komplexität der syntaktischen Konstruktionen schließ-lich ist von der Zahl der benötigten und inventarisierten Konstruktionen und der Anzahl der elementaren syntaktischen Kategorien abhängig. Dieser Zu-

[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 Konstiutenten und eine Länge von 6.

sammenhang ergibt sich aus einfacher Kombinatorik: Jede Konstruktion be-steht aus einer linearen Folge von Tochterknoten (unmittelbaren Konstituen-ten) und ist durch deren Kategorien und Reihenfolge bestimmt. Mit einer gege-benen 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.

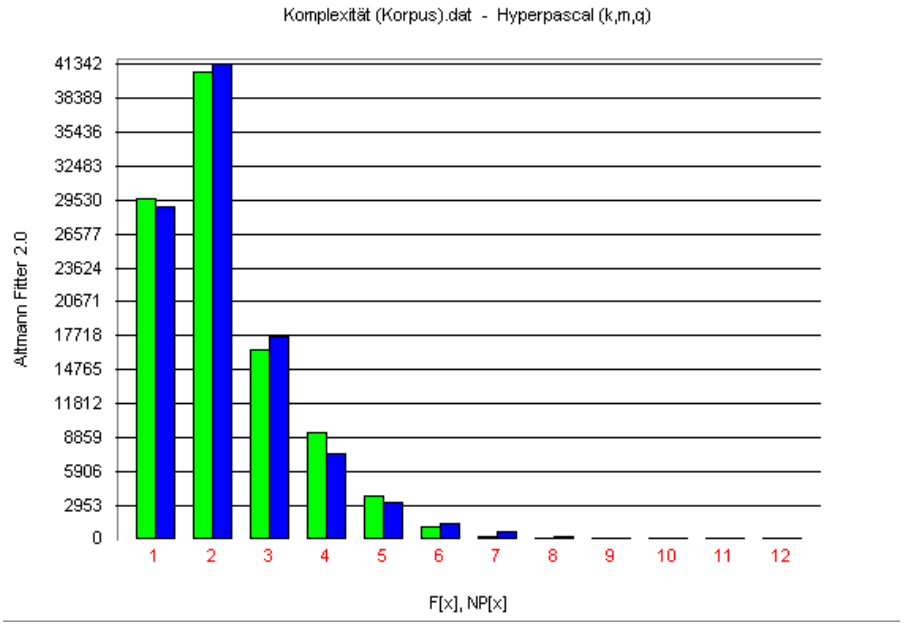


Abb. 4: Die empirische Häufigkeitsverteilung der Konstituentenkomplexität im Susanne-Korpus und Anpassung der Hyperpascal-Verteilung

Die empirische Überprüfung der Hypothesen über den Zusammenhang zwi-schen Frequenz und Komplexität und Komplexität und Länge ist in den Ab-bildungen 5, 6 und 7 dargestellt.

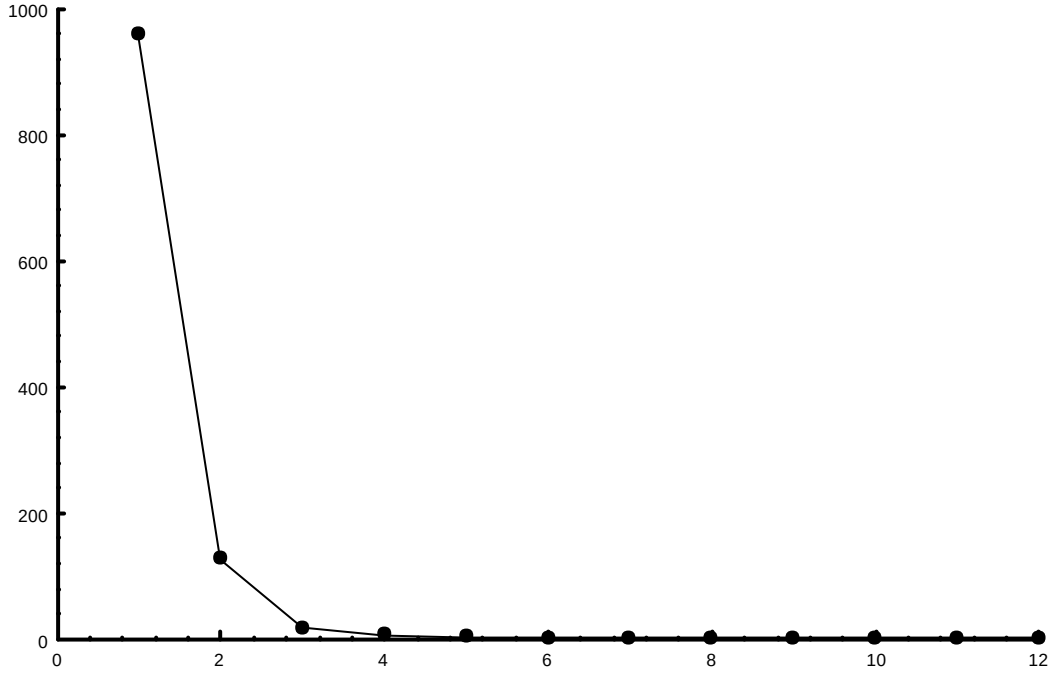


Abb. 5: Die empirische Abhängigkeit der mittleren Frequenz der Konstituenten als Funktion ihrer Komplexität. Anpassung der Funktion $F = 858{,}83\ K^{-3{,}095}e^{0{,}00727K}$ mit dem Determinationskoeffizienten D = 0,99

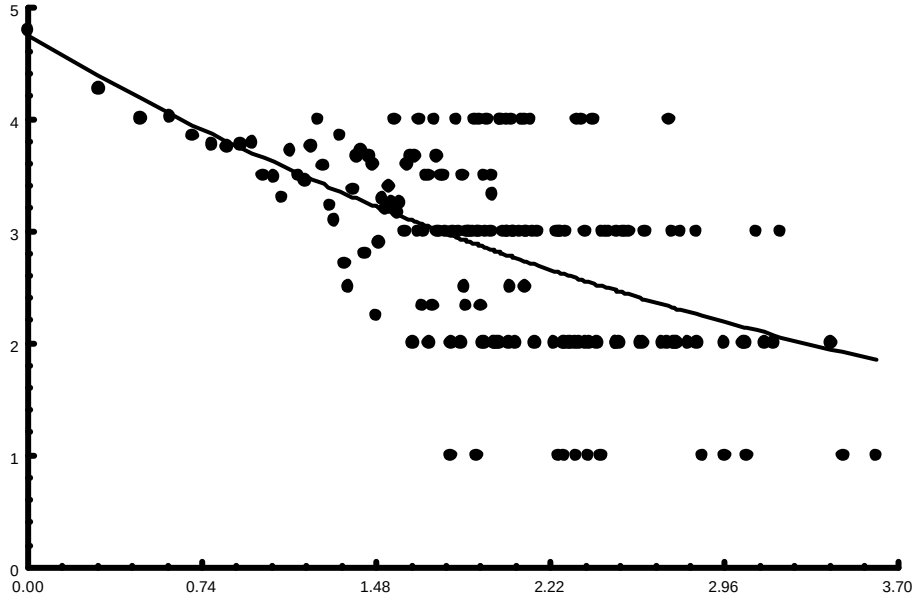


Abb. 6: Die empirische Abhängigkeit der mittleren Komplexität als Funktion der Frequenz und Anpassung der Funktion $K = 4{,}789\ F^{-0{,}1160}$ (Determinationskoeffizient D = 0,331). Die x-Achse wurde logarithmiert.

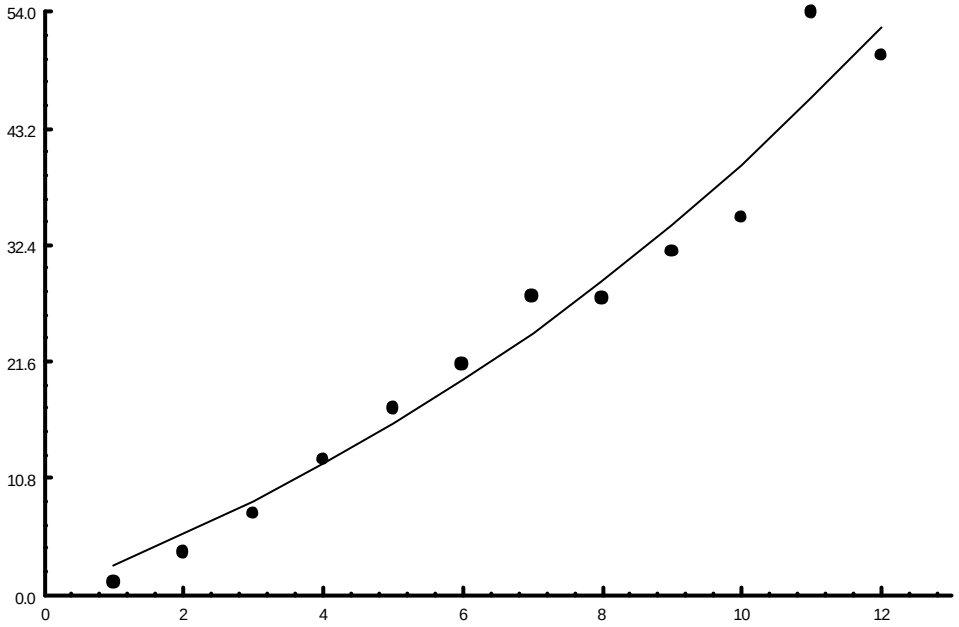


Abb. 7: Der empirische Zusammenhang zwischen Komplexität und Länge. Anpassung der aus der Modellstruktur abgeleiteten Funktion $L = 2{,}603\ K^{0{,}963} e^{0{,}0512\ K}$ mit dem Determinationskoeffizienten D = 0,960

Die vermuteten Tendenzen bestätigen sich offensichtlich. Obwohl (außer im Fall der Abhängigkeit der Länge von der Komplexität) noch keine theo-retisch begründeten Hypothesen über die mathematische Form der Zusammen-hänge vorliegen und daher kein ernsthafter Signifikanztest durchgeführt wer-den kann, muß man die generelle Hypothese über die Existenz einer inversen Abhängigkeit aufgrund der Daten als empirisch haltbar machen.

Diesen Befunden, besonders dem Frequenzspektrum, kommt potentiell auch eine wichtige praktische Bedeutung zu: Von den 4621 verschiedenen Konstituententypen mit ihren 90821 Vorkommen im gesamten 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 Gesamt-inventar) vor etc. Weniger als 20% der entsprechenden Grammatikregeln sind also öfter als viermal, weniger als 30% der Regeln öfter als zweimal anwendbar (vgl. Tab. 1).

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

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

Es ist zu erwarten, daß die Untersuchung anderer Korpora, auch anderer Sprachen als des Englischen, 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 Auslegung von Wörterbüchern (bzw. der Bestimmung ihres Textabdeckungsgrads) Verwendung finden, könn-te 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.

Länge, Komplexität und Position

Ein bereits von Otto Behaghel in (Behaghel, 1930) festgestellter Zusammen-hang 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 (Uhlírová, 1997a, 1997b). In Abbildung 8 wird dieser Zusammenhang gra-phisch in einer Form dargestellt, die seine Integration in ein umfassendes syn-taktisches Teilmodell ermöglicht. Anstelle der Länge in Wortanzahl ist hier die Komplexität als relevante Größe angesetzt, da es ja nicht um zwischen-gespeicherte Wörter, sondern um Knoten im Syntaxbaum geht. Daß sich die Wirkung des Prinzips auch an der Länge in Wörtern nachweisen läßt, ist offen-sichtlich ein indirekter Effekt.

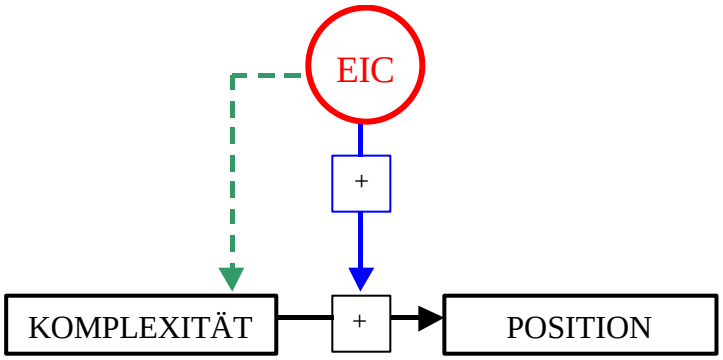


Abb. 8: Hawkin’s EIC-Prinzip modifiziert (Komplexität anstelle von Länge)

Zur nochmaligen Überprüfung dieser Hypothese wurden ebenfalls Daten aus dem Susanne-Korpus herangezogen. Abweichend von den anderen Untersuchungen wurde nicht die Stellung ‚gleichwertiger Konstituenten‘ verschie-dener Länge paarweise verglichen, sondern es wurden die Daten zu Länge, Komplexität und absoluter Position in der jeweiligen Mutterkonstituente so-wohl 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 Abhängig-keiten sind so deutlich, daß sich ein Signifikanztest erübrigt, zumal zur Zeit noch keine theoretische Begründung für die Annahme einer speziellen Funk-tion vorliegt, die mittels Anpassung an die Daten geprüft werden könnte. Bei der Entwicklung einer entsprechenden Formel werden außer Hawkins’ Über-legungen 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.

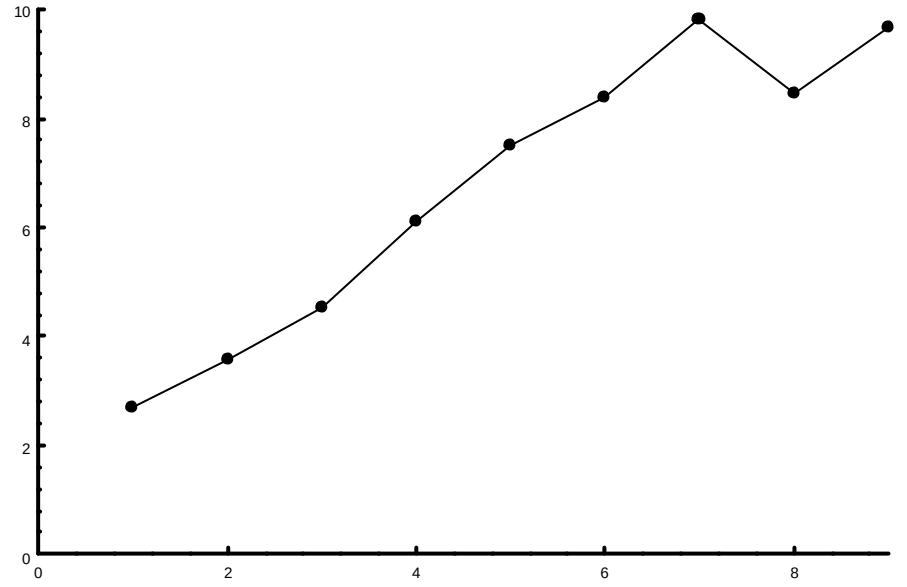


Abb. 9. Die empirische Abhängigkeit der mittleren Konstituentenlänge (in Wörtern) und der Position der jeweiligen Konstituente in der Mutterkonstituente. Die Werte für Positionen oberhalb von 8 wurden wegen mangelnder Belegzahl (<10) nicht berücksichtigt.

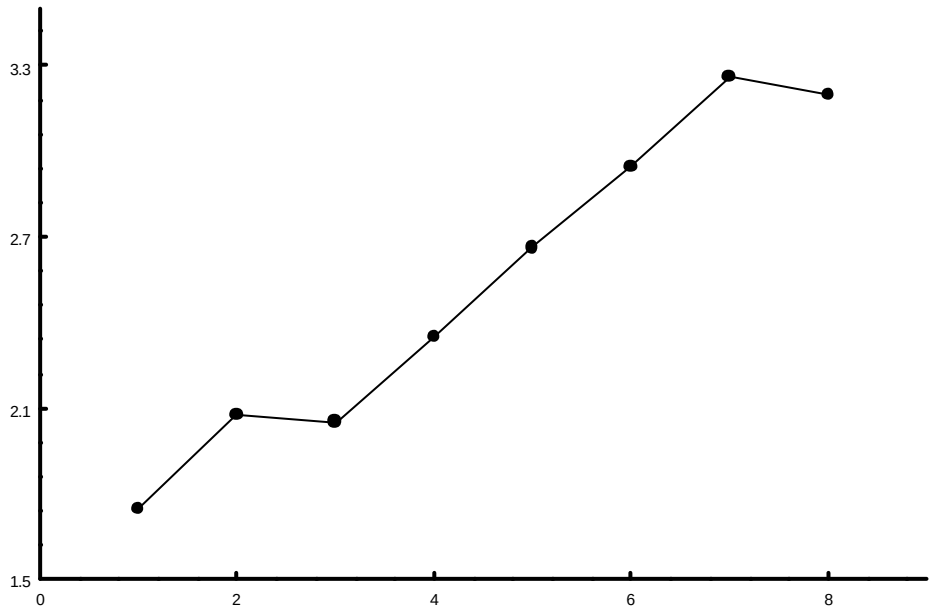


Abb. 10: Die empirische Abhängigkeit der mittleren Konstituentenkomplexität (in unmittelbaren Konstituenten) und der Position der jeweiligen Konstituente in der Mutterkonstituente. Die Werte für Positionen oberhalb von 8 wurden wegen mangelnder Belegzahl (<10) nicht berücksichtigt.

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 Verschachtelungstiefe[4] zeigen. Um dies zu überprüfen, wurden Verschachtelungstiefe (die Satzebene wurde als Tiefe 1 ge-zählt) 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. Ab-bildung 12 zeigt den Zusammenhang zwischen Verschachtelungstiefe und Posi-tion im Satz.

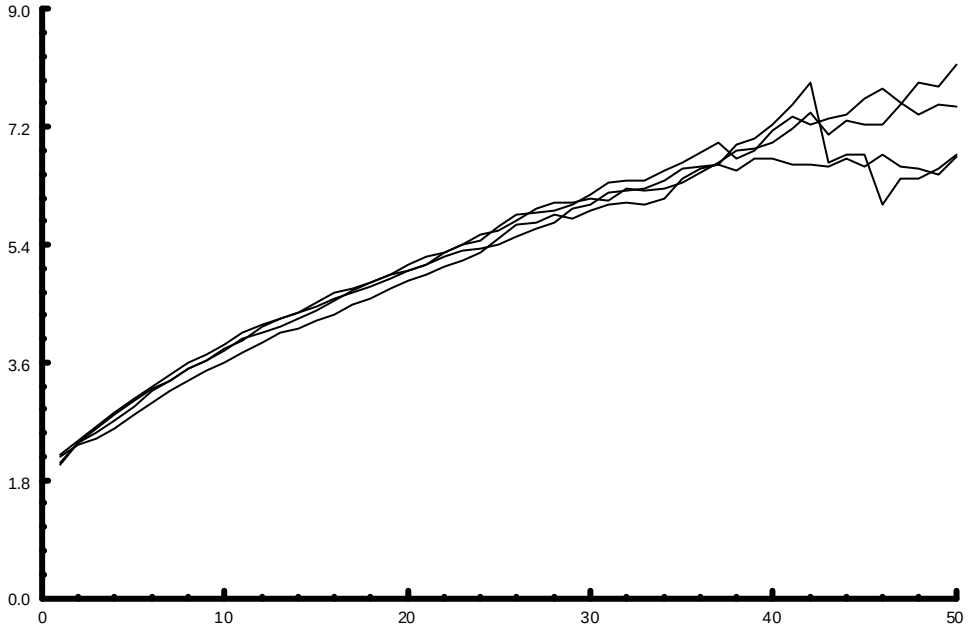


Abb. 11: Die empirische Abhängigkeit der Verschachtelungstiefe von Konstituenten von ihrer Position (gemessen in laufenden Wörtern vom Satzbeginn) für vier Teilkorpora (Texttypen) gesondert. Daten oberhalb der Wortposition 50 sind wegen geringer Belegungsdichte unzuverlässig und wurden nicht dargestellt.

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

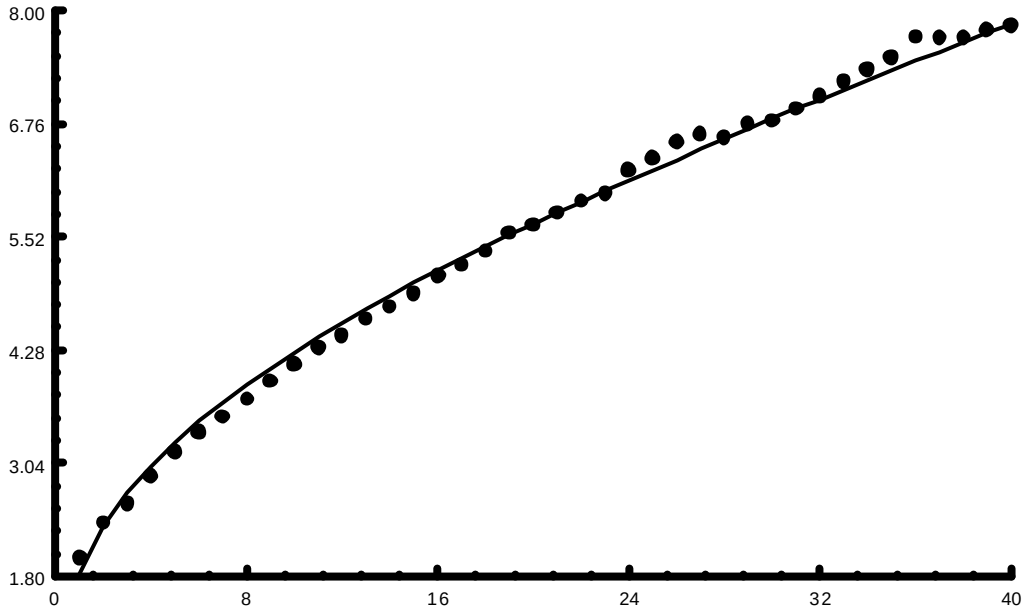


Abb. 12: Die empirische Abhängigkeit der Verschachtelungstiefe von Konstituenten von ihrer Position (gemessen in laufenden Wörtern vom Satzbeginn) für das Gesamtkorpus. Daten oberhalb der Wortposition 40 sind wegen geringer Belegungsdichte unzuverlässig und wurden nicht dargestellt. Anpassung der Funktion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

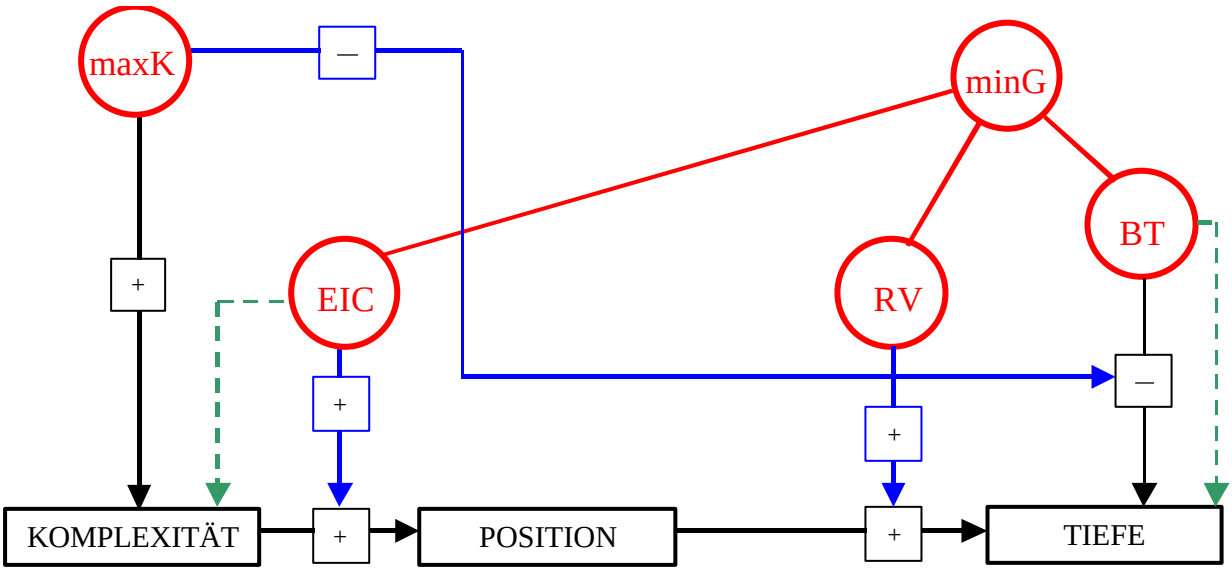


Abb. 13: Modellausschnitt mit den Größen Komplexität, Position und Tiefe und den dazugehörigen Anforderungen

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 zwischenzuspeichernden 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 tatsäch-lich gefundenen Typs erforderlich. Wenn die hier angestellte Überlegung zu-trifft, wenn also die Sprachen mit Hilfe ihrer Selbstorganisationsmechanismen ihre Grammatiken so einrichten – bzw. das Sprachverhalten dafür sorgt –, daß eine Maximierung der Komplexität der verarbeitbaren Konstituenten auf diese Weise erreicht wird, sollte der Logarithmus (als Maß für die Größe des erfor-derlichen Kodes) 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.

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 Konstitutionenten (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.

Zur Zeit 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 (s.

Köhler, 1984, wo eben dieses aus einfachen Überlegungen zu Eigenschaften der Sprachverarbeitung abgeleitet wird).

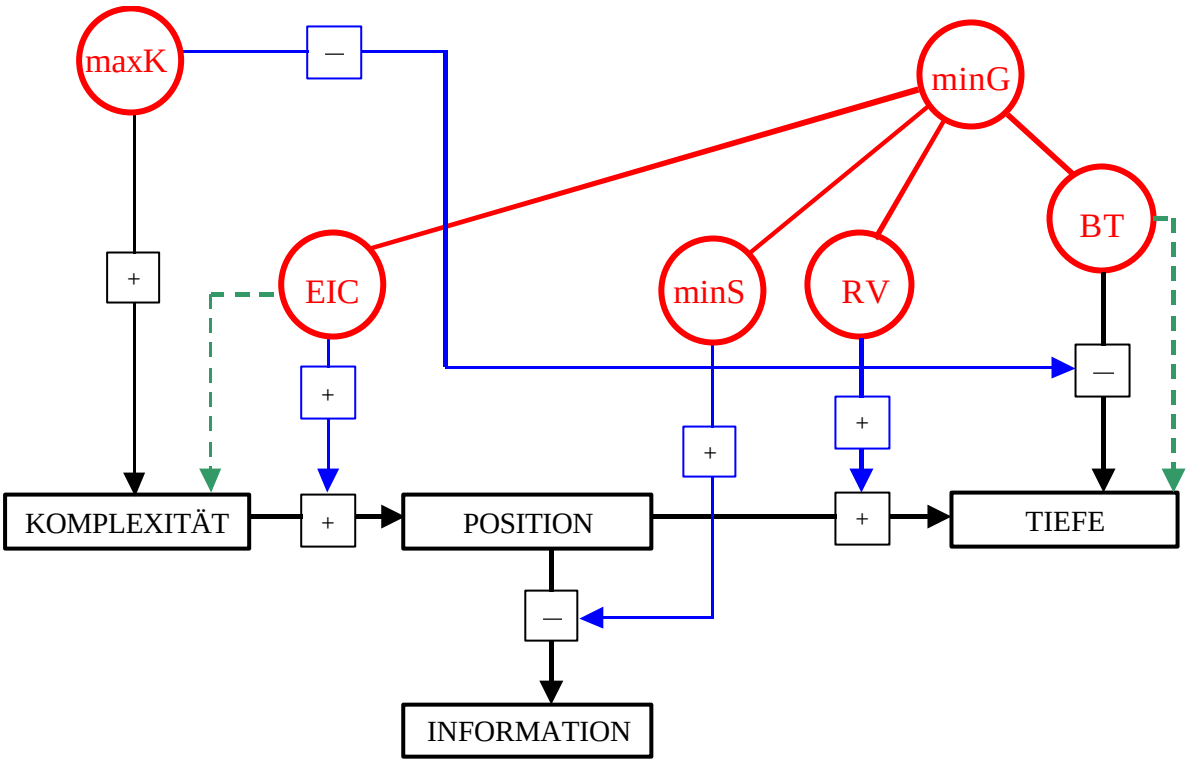


Abb. 14: Mit wachsender Position sinkt die Menge an zu speichernder Strukturinformation

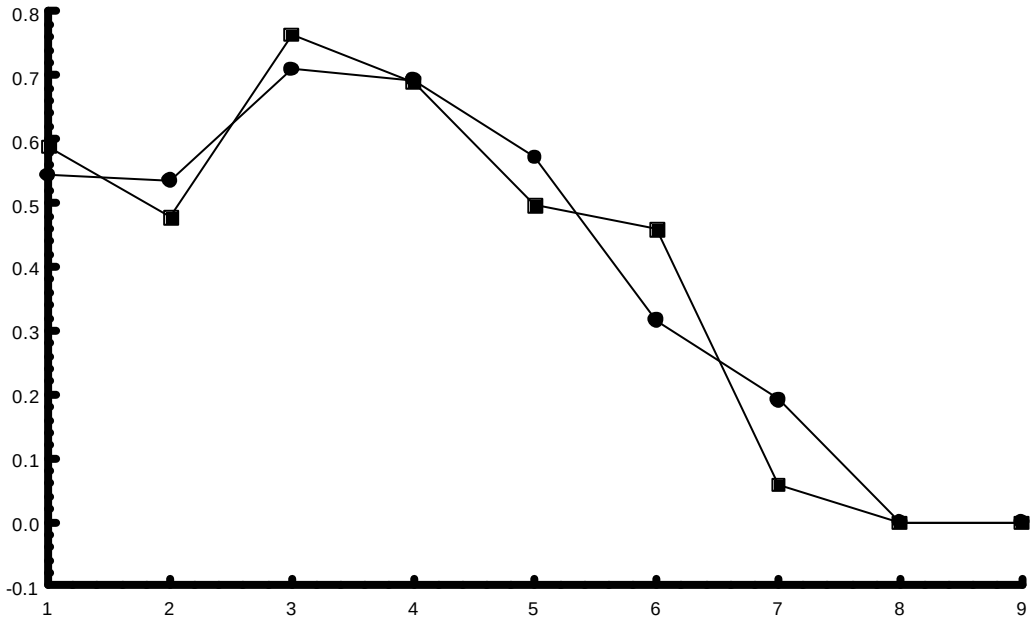


Abb. 15: Logarithmus der Anzahl alternativ möglicher Konstituententypen in Abhängigkeit von der Position (für zwei der vier Textsorten im Korpus gesondert berechnet)

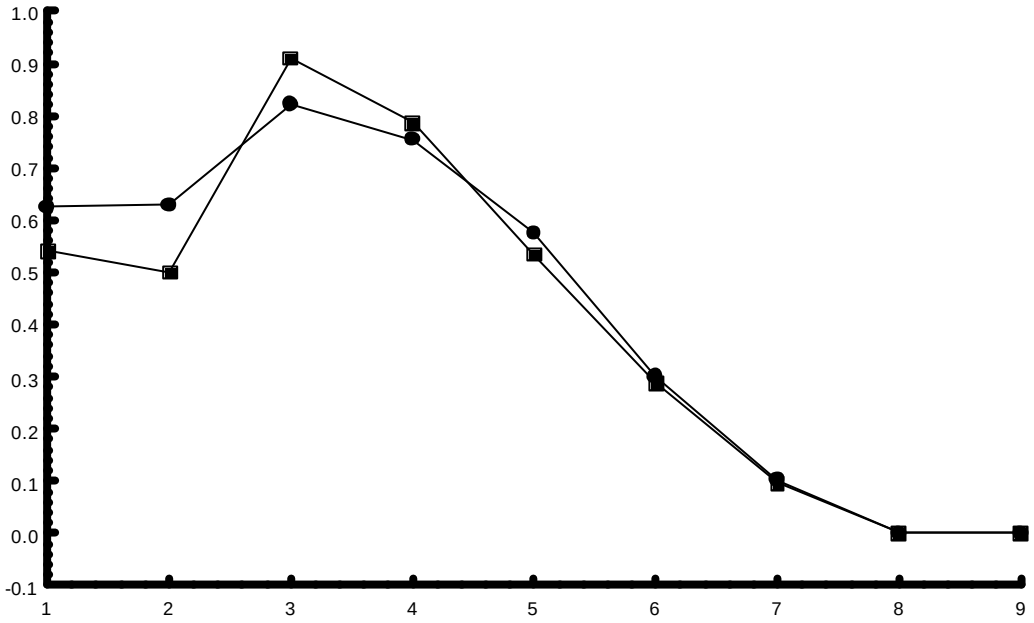


Abb. 16a: Logarithmus der Anzahl alternativ möglicher Konstituentenfunktionen in Abhängigkeit von der Position (für zwei der vier Textsorten im Korpus gesondert berechnet).

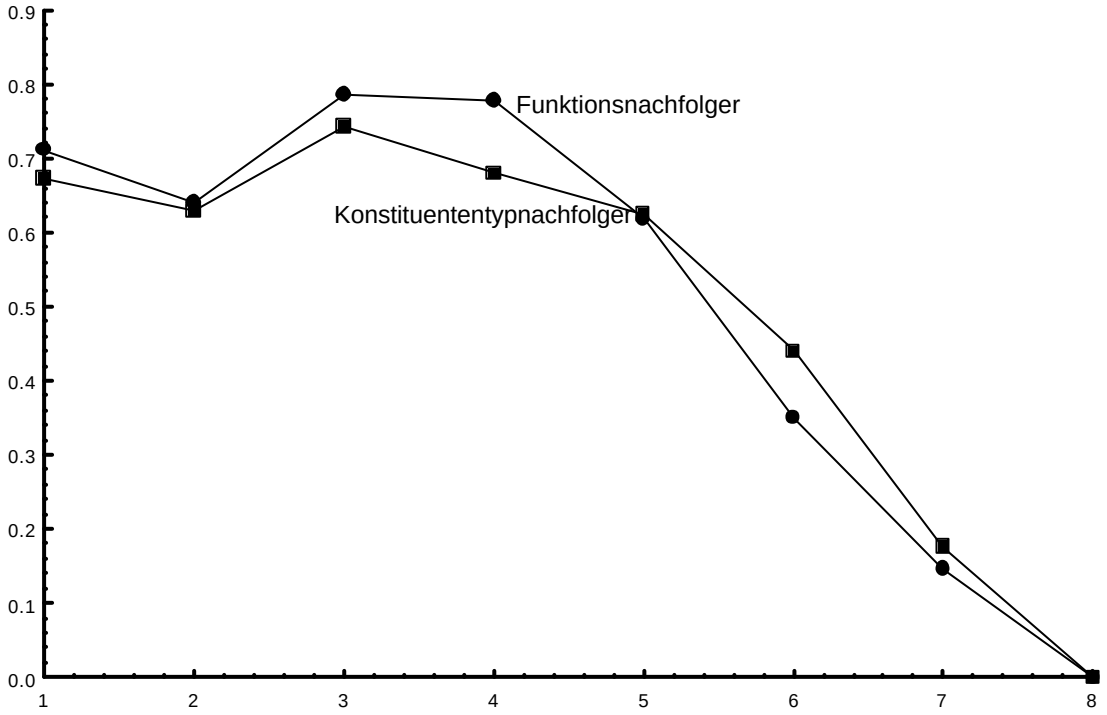


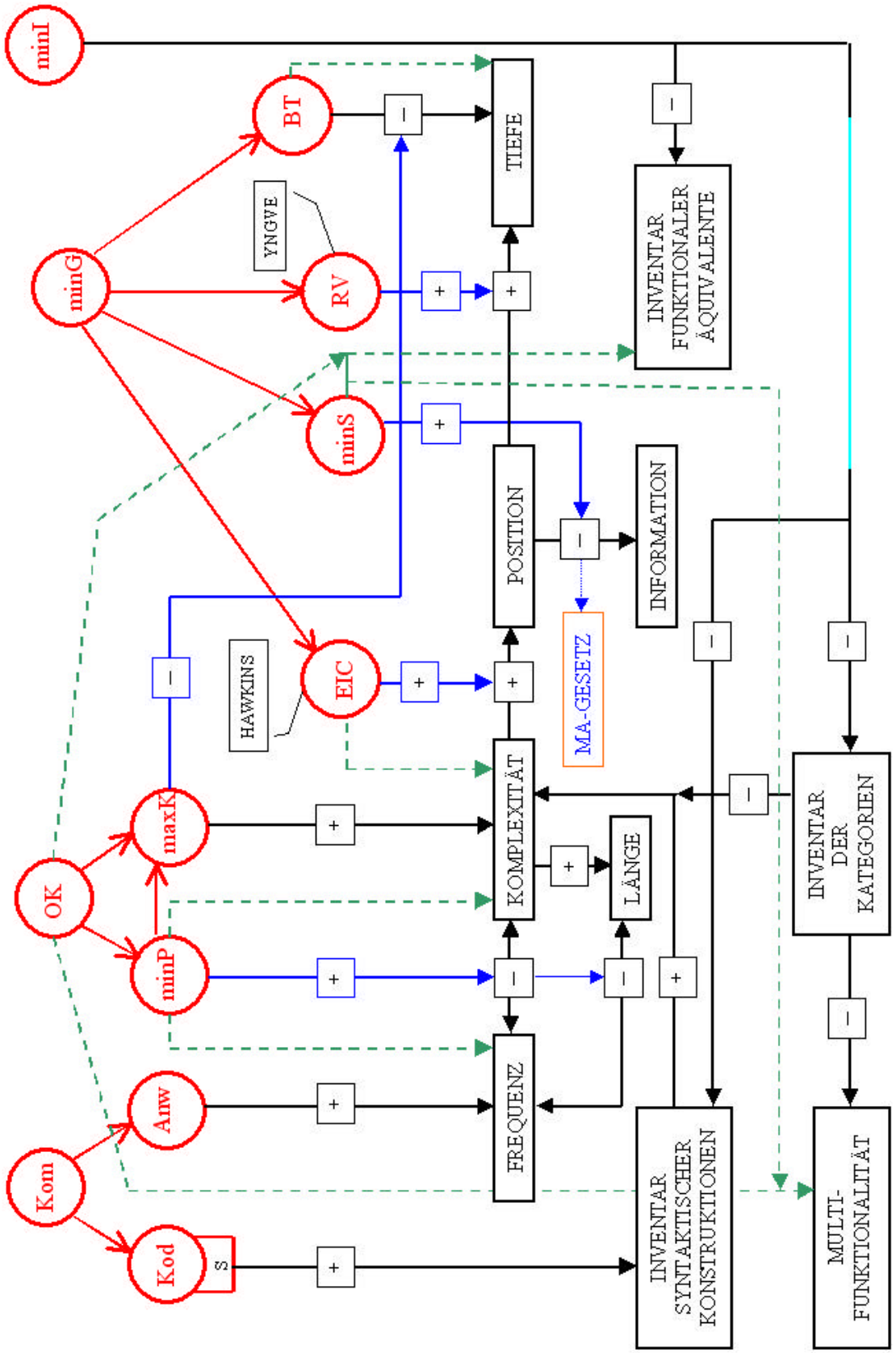
Abb. 16b: Logarithmus der Anzahl alternativ möglicher Konstituententypen bzw. Konstituentenfunktionen in Abhängigkeit von der Position (für einen einzelnen Text gesondert berechnet)

Weitere Größen und Anforderungen

Wie in anderen, früher aufgestellten Teilmodellen wird auch für das syntak-tische Subsystem ein Systembedürfnis nach Inventarminimierung (minI) pos-tuliert. 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 Haufigkeitsverteilungen innerhalb der Inventare werden durch Ordnungsparameter gesteuert (vg. Abb. 17).

Die theoretische und empirische Analyse der Wahrscheinlichkeitsverteilung der betrachteten Größen (Frequenz, Länge, Komplexität, Position, Tiefe und Information) und der Rang-Frequenz-Verteilungen der Inventare soll einer späteren Publikation vorbehalten bleiben. 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 einführend betont, lediglich ein erster Versuch, dieses Subsystem im Rahmen des synergetischen Ansatzes zu modellieren, der zur Zeit in verschiedener Hinsicht unvollständig geblieben ist. Vor allem fehlt noch die theoretische Ableitung der mathematischen Form einiger der funktionalen Zusammenhänge und der eben erwähnten Verteilungen. 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 Zusammen-hang zwischen der dargestellten Modellstruktur und dem Menzerathschen (auch Menzerath-Altmann-) Gesetz auszuarbeiten sein.


minI
BT
minG
YNGVE
RV
minS
EIC
HAWKINS
maxK
OK
minP
Anw
Kom
Kod
s
TIEFE
INVENTAR FUNKTIONALER ÄQUIVALENTE
POSITION
INFORMATION
MA-GESETZ
KOMPLEXITÄT
LÄNGE
FREQUENZ
INVENTAR DER KATEGORIEN
INVENTAR SYNTAKTISCHER KONSTRUKTIONEN
MULTI-FUNKTIONALITÄT

Literatur

Behaghel, O. (1930). Von deutscher Wortstellung. *Zeitschrift für Deutschkunde*, 44, 81-89.

Gieseking, K. (2002). Untersuchungen zur Synergetik der englischen Lexik. In diesem Band.

Hammerl, Rolf (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines Basismodells*. Trier: Wissenschaftlicher Verlag Trier.

Hawkins, J. (1994). *A performance theory of order and constituency*. Cambridge: University Press.

Hoffmann, Chr. (1996). Quantitativ-funktionalanalytische Untersuchungen zur Wortstellungsvariation. Magisterarbeit Trier.

Hoffmann, Chr. (1999). Word order and the principle of „Early immediate constituents". *Journal of Quantitative Linguistics*, 6,2,

Hoffmann, Chr., & Krott, A. (2002). Einführung in die synergetische Linguistik. In diesem Band.

Köhler, R. (1984). Zur Interpretation des Menzerathschen Gesetzes. In J. Boy & R. Köhler (Hg.), *Glottometrika 6* (S. 177-183), Trier: Wissenschaftlicher Verlag Trier.

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

Köhler, R. (1990a). Elemente der synergetischen Linguistik. In R. Hammerl (Hg.), *Glottometrika 12* (S. 179-187), Bochum: Brockmeyer.

Köhler, R. (1990b). Synergetik und sprachliche Dynamik. In W.A. Koch (Hg.), *Natürlichkeit der Sprache und Kultur* (S. 96-112), Bochum: Brockmeyer.

Köhler, R. (1991). Diversification of coding methods in grammar. In U. Rothe (Hg.), *Diversification processes in language* (S. 47-55), Hagen: Rottmann Medienverlag.

Krott, A. (1996). Some remarks on the relation between word length and morpheme length. *Journal of Quantitative Linguistics*, 3, 29-37.

Krott, A. (2002). Ein funktionalanalytisches Modell der Wortbildung. In diesem Band.

Sampson, G. (1995). *English for the Computer*. Oxford.

Uhlirová, L. (1997). Length vs. Order. Worder Length and Clause Length from the Perspective of Word Order. *Journal of Quantitative Linguistics*, 4, 266-275.

Uhlirová, L. (1997b). O vztahu mezi délkou slova a jeho polohou ve vete. *Slovo a slovesnost*, 58, 174-184.

Yngve, V. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104, 444-466.

Anhang

Das der Untersuchung zugrunde liegende Textkorpus ist das „Susanne-Korpus“ (Sampson, 1995), eine Zusammenstellung von 64 englischen – syntaktisch annotierten – Texten mit insgesamt 128000 laufenden Wörtern. Zur Illustration ist unten der erste Satz des Texts A01 dargestellt. Die organisatorische und linguistische Information wird wortweise in sechs Spalten angegeben: Die erste Spalte („reference field“) liefert einen Text- und Spaltenkode, die zweite („status field“) markiert Abkürzungen, Symbole und Schreibfehler, die dritte enthält die Wortart nach dem „Lancaster tagset“, die vierte die Wortform des Rohtexts, die fünfte das Lemma und die sechste das Ergebnis der syntaktischen Analyse. In den Zeilen A01:0040j und A01:0050d z.B. markieren die,:0’s die NP „the over-all... of the election“ als logisches direktes Objekt, die eckigen Klammern mit dem Etikett Fr in den Zeilen A01:0060h und A01:0060n bedeuten, daß „in which... was conducted“ ein Relativsatz ist.

A01:0010a	-	YB	<minbrk>	-	[Oh.Oh]
A01:0010b	-	AT	The	the	[O[S[Nns:s.
A01:0010c	-	NP1s	Fulton	Fulton	[Nns.
A01:0010d	-	NNL1cb	County	county	.Nns]
A01:0010e	-	JJ	Grand	grand	.
A01:0010f	-	NN1c	Jury	jury	.Nns:s]
A01:0010g	-	VVDv	said	say	[Vd.Vd]
A01:0010h	-	NPD1	Friday	Friday	[Nns:t.Nns:t]
A01:0010i	-	AT1	an	an	[Fn:o[Ns:s.
A01:0010j	-	NN1n	investigation	investigation	.
A01:0020a	-	IO	of	of	[Po.
A01:0020b	-	NP1t	Atlanta	Atlanta	[Ns[G[Nns.Nns]
A01:0020c	-	GG	+<apos>s	-	.G]
A01:0020d	-	JJ	recent	recent	.
A01:0020e	-	JJ	primary	primary	.
A01:0020f	-	NN1n	election	election	.Ns]Po]Ns:s]
A01:0020g	-	VVDv	produced	produce	[Vd.Vd]
A01:0020h	-	YIL	<ldquo>	-	.
A01:0020i	-	ATn	+no	no	[Ns:o.
A01:0020j	-	NN1u	evidence	evidence	.
A01:0020k	-	YIR	+<rdquo>	-	.
A01:0020m	-	CST	that	that	[Fn.
A01:0030a	-	DDy	any	any	[Np:s.
A01:0030b	-	NN2	irregularities	irregularity	.Np:s]
A01:0030c	-	VVDv	took	take	[Vd.Vd]
A01:0030d	-	NNL1c	place	place	[Ns:o.Ns:o]Fn]Ns:o]Fn:o]S]
A01:0030e	-	YF	+.	-	.O]

A01:0030f	-	YB	<minbrk>	-	[Oh.Oh]
A01:0030g	-	AT	The	the	[O[S[Ns:s.
A01:0030h	-	NN1c	jury	jury	.Ns:s]
A01:0030i	-	RRR	further	far	[R:c.R:c]
A01:0030j	-	VVDv	said	say	[Vd.Vd]
A01:0030k	-	II	in	in	[P:p.
A01:0030m	-	NNT1c	term	term	[Np[Ns.
A01:0030n	-	YH	+<hyphen>	-	.
A01:0030p	-	NN1c	+end	end	.Ns]
A01:0040a	-	NN2	presentments	presentment	.Np]P:p]
A01:0040b	-	CST	that	that	[Fn:o.
A01:0040c	-	AT	the	the	[Nns:s101.
A01:0040d	-	NNL1c	City	city	.
A01:0040e	-	JB	Executive	executive	.
A01:0040f	-	NNJ1c	Committee	committee	.
A01:0040g	-	YC	+,	-	.
A01:0040h	-	DDQr	which	which	[Fr[Dq:s101.Dq:s101]
A01:0040i	-	VHD	had	have	[Vd.Vd]
A01:0040j	-	JB	over<hyphen>all	overall	[Ns**:o**.
A01:0050a	-	NN1n	charge	charge	.
A01:0050b	-	IO	of	of	[Po.
A01:0050c	-	AT	the	the	[Ns.
A01:0050d	-	NN1n	election	election	.Ns]Po]Ns**:o**]
A01:0050e	-	YC	+,	-	.Fr]Nns:s101]
A01:0050f	-	YIL	<ldquo>	-	.
A01:0050g	-	VVZv	+deserves	deserve	[Vz.Vz]
A01:0050h	-	AT	the	the	[N:o.
A01:0050i	-	NN1u	praise	praise	[NN1n&.
A01:0050j	-	CC	and	and	[NN2+.
A01:0050k	-	NN2	thanks	thank	.NN2+]NN1n&]
A01:0050m	-	IO	of	of	[Po.
A01:0050n	-	AT	the	the	[Nns.
A01:0060a	-	NNL1c	City	city	.
A01:0060b	-	IO	of	of	[Po.
A01:0060c	-	NP1t	Atlanta	Atlanta	[Nns.Nns]Po]Nns]Po]N:o]
A01:0060d	-	YIR	+<rdquo>	-	.
A01:0060e	-	IF	for	for	[P:r.
A01:0060f	-	AT	the	the	[Ns:103.
A01:0060g	-	NN1c	manner	manner	.
A01:0060h	-	II	in	in	**[Fr**[Pq:h.
A01:0060i	-	DDQr	which	which	[Dq:103.Dq:103]Pq:h]
A01:0060j	-	AT	the	the	[Ns:S.
A01:0060k	-	NN1n	election	election	.Ns:S]
A01:0060m	-	VBDZ	was	be	[Vsp.
A01:0060n	-	VVNv	conducted	conduct	.Vsp]**Fr]**Ns:103]P:r]Fn:o]S]
A01:0060p	-	YF	+.	-	.O]