EMPIRISCHE TEXTSEMIOTIK UND QUANTITATIVE TEXT-TYPOLOGIE

PETER GRZYBEK, EMMERICH KELIH, ERNST STADLOBER

Zusammenfassung: *Die vorliegende Untersuchung zielt auf die Frage nach einer auf der Quantifizierung von Textstrukturen basierenden Text-Klassifikation. In einer text-theoretischen Einleitung wird zunächst die Problematik der Kategorisierung bzw. Klassifizierung von Texten diskutiert; dabei geht es in diesem ersten Schritt um die Fundierung eines operationalen Textbegriffes, mit dem Texte einer quantitativen Untersuchung zugänglich gemacht werden können. In einem zweiten Schritt wird die Wortlänge - unter abermaligem Einschluß der Diskussion um die mit einer Definition des Wortes verbundenen Problematik - als zentrales Beschreibungsmerkmal von Texten behandelt. Im letzten Schritt schließlich werden unterschiedliche Methoden der quantitativen Textklassifizierung (Cluster-Verfahren, post-hoc-Untersuchungen, multivariate Diskriminanzanalyse) detailliert dargelegt und angewendet. Die im Anschluß daran im einzelnen dargelegten Resultate der quantitativen Textklassifizierung, basierend auf 398 slowenischen Texten, geben einen Einblick in die Effektivität der angewandten Methoden; sie legen zudem die Notwendigkeit und Möglichkeit einer neuen Text-Typologie nahe.*

Summary: *The present study concerns the question of text classification, based on a quantitative approach to text structures. Subsequent to an introductory discussion on general aspects of text theory, the first step concentrates on an operative definition of 'text'. In a second step, word length - the definition of 'word' also being submitted to a theoretical discussion - is presented as a central characteristic of texts. Based on these theoretical foundations, various quantitative methods of text classification (such as cluster analyses, post-hoc procedures, multivariate discriminant analyses) are presented and applied to 398 Slovenian texts. Finally, the results of the quantitative text classification are presented in detail; these results do not only prove the efficiency of the method(s) presented, but also lead to a new typology of texts.*

DAS TEXT-UNIVERSUM UND SEINE STRUKTURIERUNG: TEXT-THEORETISCHE GRUNDÜBERLEGUNGEN

Die Kategorisierung von Texten ist eines der ältesten Anliegen der Sprach- und Literaturwissenschaften, angefangen von Fragen der Gattungstheorie über verschiedene Versionen der Stilistik bis hin zur gegenwärtigen Textsortenforschung. Während es in der Literaturwissenschaft bei der Bestimmung und Unterscheidung von Gattungen lange Zeit um die Gegenüberstellung von formalen und inhaltlichen Kriterien ging, bildete sich Anfang des 20. Jahrhunderts mit dem Russischen Formalismus, nicht zuletzt unter dem Einfluß der sich herausbildenden synchronen Sprachwissenschaft, eine funktionale Sichtweise heraus; diese stellte zunächst textinterne funktionale Aspekte in den Vordergrund

der Betrachtung und versuchte dann, diese mit textexternen (pragmatischen, soziologischen o.a.) Faktoren in Beziehung zu setzen. So wurden insbesondere im Umfeld des Tschechoslowakischen Strukturalismus der 30er und 40er Jahre die Grundlagen der sog. Funktionalstilistik gelegt, der es darum ging, Fragen der Stilistik nicht nur im individuellen, sondern auch inter-individuellen Bereich zu verankern, und der es dabei daran gelegen war, stilistische Merkmale auf der textinternen Ebene mit pragmatischen Funktionen auf der textexternen Ebene zu verbinden. In der weiteren Folge ist man davon ausgegangen, daß die Kombination einer Reihe obligatorischer bzw. fakultativer Stilzüge für einen bestimmten Funktionalstil charakteristisch ist, und daß andererseits ein Funktionalstil einer charakteristischen gesellschaftlichen Funktion entspricht. So sind einige Interpreten der Funktionalstilistik wie z.B. Scharnhorst (1981: 305) sogar so weit gegangen, den Grundgedanken der Funktionalstilistik darin zu sehen, daß „verschiedene Typen gesellschaftlicher Tätigkeit unterschiedliche Arten sprachlich-kommunikativer Tätigkeit und dementsprechend eine Differenzierung des sprachlichen Zeichensystems erfordern."

In dieser Tradition steht letztendlich auch die aktuelle Textsortenforschung, insofern sie Textsorten als Klassen von Texten versteht, die von gemeinsamen inhaltlichen („thematisch-propositionalen"), stilistischen und handlungs-typisch illokutiven Grundelementen bestimmt sind. Trotz dieses den beiden Zugängen gemeinsamen Interesses an einer spezifischen Korrelation pragmatischer und stilistischer Komponenten befinden sie sich im Rahmen pragmatisch ausgerichteter Ansätze allerdings in gewisser Weise an zwei unterschiedlichen Enden des Spektrums: Denn der bedingt als top-down-orientiert zu bezeichnenden Funktionalstilistik war stets an einer maximalen Reduktion (d.h. an der Bestimmung einer minimalen Anzahl) von Textklassen gelegen, während es der Textsortenforschung um eine maximale Ausdifferenzierung des textkommunikativen Geschehens geht.

Der Unterschied zwischen beiden Herangehensweisen liegt letzten Endes somit darin, daß der Differenzierung von Funktionalstilen allgemeine kommunikative (gesellschaftlich definierte) Sprach- und/oder Textfunktionen zugrundegelegt werden, während es sich bei der Textsortenforschung um spezifische kommunikativ-situative Funktionen handelt. Mit dieser unterschiedlichen Zugangsweise ist natürlich eine extreme Unterschiedlichkeit in der Anzahl der resultierenden Kategorien verbunden: So bringt die Textsortenforschung es mittlerweile auf ein Inventar von nicht weniger als ca. 4000 verschiedenen Textsorten (vgl. Adamzik 1995: 255ff.). Allein innerhalb des Texttyps ‚Brief' werden mehrere Dutzend Sorten unterschieden, angefangen von Ablaß- und Abschiedsbriefen über Beileids- und Bittbriefe bis hin zu Zulassungsbriefen.

Im Vergleich dazu hat die Funktionalstilistik sich – je nach konkreter Schule – in der Regel mit der Unterscheidung von ca. fünf bis acht verschiedenen Funktionalstilen begnügt, die – zumindest im Bereich der Stilistik – als die höchsten und abstraktesten Kategorien angesehen werden. Allerdings wurden in dieser idealtypologischen Differenzierung Überlappungs- und Übergangsbereiche zwischen den Funktionalstilen nicht erfaßt; so blieb selbst Proponenten der Funktionalstilistik schließlich und letzten Endes nicht Anderes übrig als einzugestehen, daß sie eine Antwort auf die Frage schuldig bleiben mußten, „*welche* funktionalen Stile/Substile und *wieviele* objektiv nachgewiesen werden können" (Riesel/Schendels 1975: 19).

Nicht zuletzt dieser Umstand wird dafür mitverantwortlich gemacht, daß die Funkti-

onalstilistik in der Diskussion der letzten zehn Jahre gegenüber textlinguistischen und pragmatischen Überlegungen in den Hintergrund getreten zu sein scheint (Blühdorn 1990: 218): „Anstatt für die großen Stilbereiche interessiert man sich gegenwärtig mehr für kleinere Einheiten – sei es die linguistische Charakteristik von Einzeltexten [...] oder überschaubaren Textsorten [...]." Letztendlich ist damit das Problem jedoch weit von einer Lösung entfernt: Denn wenn auch diese Einschätzung zweifellos zutrifft, wäre es überaus angebracht, über diese nüchterne Feststellung hinausgehend auch der Ursache dieses Umstands nachzugehen und es nicht bei dem simplen Verweis auf fehlende Ansatzpunkte zur Charakterisierung von (Übergängen zwischen) Funktionalstilen zu belassen (Blühdorn 1990: 218). Das tatsächliche Problem scheint vielmehr darin zu liegen, daß die Funktionalstilistik (und in deren Folge dann auch die Textsortenforschung) bei aller Unterschiedlichkeit ein entscheidendes Merkmal gemeinsam haben: Sie sind nahezu ausschließlich qualitativ ausgerichtet. Das heißt, daß in beiden Fällen die „Welt der Texte" mit Bezug zur Welt strukturiert wird, insofern die Text-Welt unter Heranziehung von textexternen (pragmatischen) Faktoren strukturiert werden soll. Solange jedoch keine operationalen Kriterien entwickelt werden, die in weiterer Folge eine empirisch fundierte Quantifizierung erlauben, wird sich jedwede Art von Text-Typologie im Kreise drehen.

In diesem Sinne waren Ansätze wie etwa der Versuch von Mistrík (1973), eine *Exakte Typologie von Texten* zu erarbeiten, durchaus ein Schritt in die richtige Richtung: Es ging darum, durch die Bezugnahme auf exakte (mathematische) Methoden Texteigenschaften meßbar zu machen, um so eine objektivierbare Basis einer Texttypologisierung bereitzustellen. Zu diesem Zweck war es zunächst notwendig, zwischen Individualstilen und Interindividualstilen zu unterscheiden. Denn die anfangs wesentlich vom innersprachlichen Systembegriff Saussures geprägte Funktionalstilistik war zunächst zwangsläufig gezwungen, funktionalstilistische Differenzierungen dem Bereich der *parole* zuzuordnen; so verstand Havránek (1942: 155) etwa Stil als die „individualisierende (spezifische) Organisation eines sprachlichen Strukturkomplexes, d.h. jeder gegebenen sprachlichen Äußerung". Erst sehr viel später wurde diese Dichotomisierung zu Recht in Zweifel gezogen, und zwar mit dem Argument, daß „die Beherrschung der funktionalstilistischen Differenzierung auf interindividuellen Regeln beruht, die in die Kompetenz des Sprechers eingehen" (Steube 1974: 115).

Vor diesem Hintergrund lag es nahe, durchaus individualstilistischen Faktoren Raum zu lassen, daneben aber Interindividualstile zu untersuchen, unter denen eben die „traditionellen" Funktionalstile, wie sie etwa in *Abb. 1* dargestellt sind, zu verstehen sind (vgl. Mistrík 1973: 23ff.).

Ungeachtet des im Prinzip richtigen Strebens nach Exaktheit der Typologie war das damalige Vorgehen jedoch mit einer Reihe von Schwächen verbunden, die de facto sehr viel mehr für das Scheitern der Funktionalstilistik und die Favorisierung detaillierterer Texttypologien verantwortlich sind als die oben erwähnte fehlende Berücksichtigung von Überlappungen und Überschneidungen zwischen den Stilen: So wurde vor allem die aufgestellte Typologie nicht wirklich in den empirischen Untersuchungen zur Disposition gestellt, sondern es wurden lediglich innerhalb der (im Prinzip uneingeschränkt akzeptierten) Funktionalstile die jeweiligen Stileigenschaften empirisch untersucht; dies entspricht letztendlich einer „blinden" Akzeptanz der Funktionalstile und versperrt die Möglichkeit einer quantitativ basierten Typologie den Weg. Weiterhin wurden individuelle Texte (z.T.

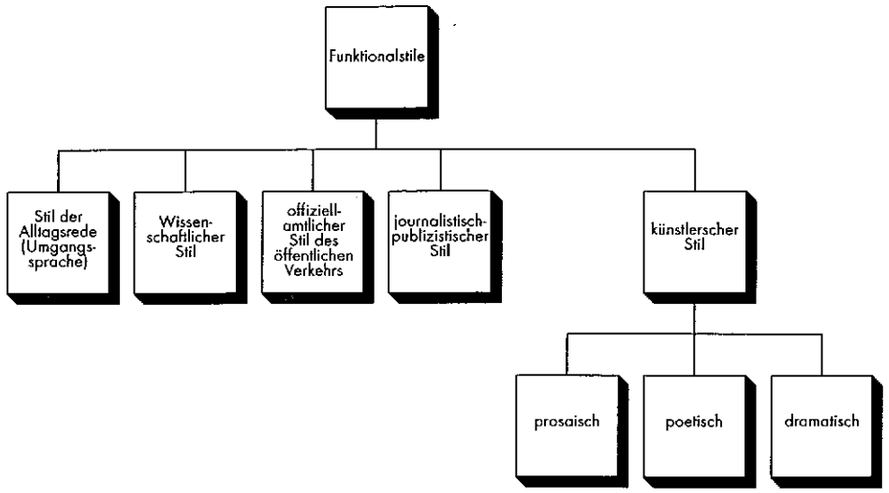


Abb 1: Funktionalstile nach Mistrik (1973)

implizit und unreflektiert, auf jeden Fall aber subjektiv-autoritativ) jeweiligen Funktionalstilen zugeordnet, was – wie oben bereits im Zusammenhang mit der Untersuchung von Grzybek & Kelih (2005b) erwähnt – sogar auf der höheren Ebene der Textsorten nicht unproblematisch ist. Und schließlich wurden bei der Untersuchung oft mehrere Texte miteinander kombiniert und zu kleinen, funktionalstilistisch differenzierten Korpora zusammengeführt; dies geschah natürlich in der Absicht, möglichst „repräsentatives" Textmaterial zu untersuchen; doch erhöht sich dabei zum einen die Gefahr der subjektiven Zuordnung der Texte zu den Funktionalstilen, zum anderen werden textinterne Stil-Spezifika relativiert und gegebenenfalls allein durch die Kombination getilgt.

Ungeachtet dessen ist eines der Grundanliegen der Funktionalstilistik nach wie vor von aktuellstem Interesse für die Textforschung: Sieht man einmal von den in der Geschichte der Funktionalstilistik mitunter vorgenommenen Korrelationen stilistischer mit soziologischen (und nicht immer ideologiefreien) Faktoren ab, ist nämlich die Frage nach einem Minimalset objektivierbarer Texttypen auf der höchsten und abstraktesten Ebene der Textklassifikation nach wie vor offen.

Vor diesem Hintergrund liegt es nahe, die im Prinzip richtig gestellten Fragen unter Vermeidung der oben angeführten methodologischen Mängel neu zu stellen.

DER ‚TEXT' AUS EMPIRISCHER SICHT

Abgesehen von dieser im text-theoretischen Bereich sehr unterschiedlichen Ausgangslage ist anzunehmen, daß wir es im Objektbereich zunächst einmal mit einem Universum von Texten zu tun haben, d.h. mit einer endlichen oder unendlichen Anzahl textueller Objekte, die ein offenes oder geschlossenes System verkörpern. Als erstes stellt sich deshalb die

prinzipielle Frage, ob dieses Universum strukturiert ist bzw. strukturiert werden kann, und wenn ja: wie. Wenn in weiterer Folge eine solche Struktur existiert – und davon ist auszugehen –, dann beinhaltet eine Beschreibung des Text-Universums zwei verschiedene Prozesse, die notwendigerweise ineinander greifen müssen:

(a) die Identifikation der Objekte des Systems, die auf einer Definition von ‚Text' beruht,
(b) die Klassifikation dieser Objekte, die in der Identifikation und Beschreibung hierarchisch geordneter Subsysteme resultiert.

Unter Beibehaltung der astronomischen Analogie – die keineswegs nur als Bild oder Metapher, sondern als wissenschaftstheoretisches Postulat zu verstehen ist – kommt es als nächstes also darauf an, innerhalb des Text-Universums möglicherweise existierende (Text-) Galaxien zu identifizieren, welche Attraktoren für die individuellen (Text-)Objekte darstellen. Schließlich gilt es, innerhalb solcher Galaxien spezifische Sub-Systeme niederer Ordnung zu identifizieren, die in Analogie zu Stern- oder Sonnensystemen zu sehen sind. Es ist selbstverständlich, daß die beiden Prozesse der Identifikation und Klassifikation nicht ohne Rückgriff auf bestimmte (text-)theoretische Annahmen realisiert werden können, welche obligatorische und/oder fakultative Merkmale der zur Diskussion stehenden Objekte betreffen. Denn den Objekten selbst sind weder quantitative noch qualitative Eigenschaften immanent; vielmehr erweisen sich letztere als das Ergebnis spezifischer kognitiver Prozesse.

Dabei involviert *jede* Art von Klassifikation auf die eine oder andere Art und Weise und in unterschiedlichem Maße quantifizierende Vorgangsweisen; der Grad der Quantifizierung wird lediglich von den in der jeweiligen Meta-Sprache enthaltenen Eigenschaften bestimmt. Deswegen ist es von besonderer Bedeutung, von welcher analytischen Ebene der Klassifikationsprozess ausgeht, wobei jede Ebene mit jeweils spezifischen Problemen behaftet ist, was die Definition der Sub-Systeme und deren Grenzen betrifft.

Ungeachtet der oben angesprochenen Probleme läßt sich die Klassifikation des Text-Universums nicht ohne empirisch-quantitative Methoden erreichen. An dieser Stelle kommt die quantitative Linguistik als eine wichtige sprachwissenschaftliche Disziplin ins Spiel: im Gegensatz zu den oben beschriebenen sprach- und textwissenschaftlichen Richtungen strebt die quantitative Linguistik nach der Entdeckung von Regularitäten und Gesetzmäßigkeiten im System der Sprache, abzielend auf eine empirisch fundierte Theorie der Sprache. Die Transformation beobachteter sprachlicher Daten in Quantitäten wird hier als ein standardisierter Zugang zur Beobachtung verstanden. Spezifische Hypothesen werden statistisch getestet; im Idealfall wird die Interpretation der Ergebnisse in einen theoretischen Rahmen eingebettet.

Eine derartige Vorgehensweise, die sich als quantitative Textanalyse bezeichnen läßt, ist durch den folgenden Grundgedanken gekennzeichnet: Im Gegensatz zur Korpuslinguistik richtet sich die quantitative Textanalyse auf individuelle Texte als homogene Einheiten. Der Unterschied ist von entscheidender Bedeutung: Geht es der Korpuslinguistik als einer primär empirisch ausgerichteten Disziplin darum, möglichst viele Texte im Hinblick auf eine angenommene Repräsentativität zum Zwecke der Untersuchung zusammenzuführen, lautet eine zentrale Grundannahme der quantitativen Textanalyse, daß ein (vollständiger) Text ein selbst-regulierendes System ist, das von spezifischen Regularitäten geregelt wird. Diese Regularitäten müssen nicht zwangsläufig auch in Textsegmenten wir-

ken, und es kann als wahrscheinlich angesehen werden, daß diese steuernden Regulationsmechanismen sich in jeglicher Art von Textkombination überschneiden. Damit ist natürlich noch keine Definition von ‚Text' geliefert, und es bleibt zu fragen, was ein ‚Text' ist: ein vollständiger Roman, das aus mehreren Kapiteln bestehende Buch eines Romans, ein einzelnes Kapitel, oder womöglich sogar einzelne Absätze, dialogische oder narrative Sequenzen? In letzter Konsequenz gibt es keine a priori verfügbare Antwort auf dieses Grundproblem in den verschiedenen Textwissenschaften, und nicht zuletzt deshalb taucht die Frage, ob es eines „neuen" Textbegriffs bedarf, mit schöner Regelmäßigkeit immer wieder in den entsprechenden textwissenschaftlichen Debatten auf. In der vorliegenden Darstellung kann es freilich nicht um eine theoretische Lösung dieser Frage gehen.

Aus unserer Sicht, die einen empirisch geprägten Blick auf das Problem wirft, stellt sich der Sachverhalt in Form von zweierlei Problemen dar: (i) das Problem der Datenhomogenität, und (ii) das Problem der zugrundegelegten Analyseeinheit(en). Aus dieser Perspektive müssen zwei spezifische Entscheidungen getroffen werden, welche die Rahmenbedingungen unserer Untersuchung repräsentieren:

1. Wir betrachten einen Text als das Resultat eines homogenen Prozesses der Textgenerierung; deshalb konzentrieren wir uns u.a. auf Briefe, Zeitungskommentare oder Kapitel von Romanen als individuelle ‚Texte'. Ausgehend von der Annahme, daß ein solcher ‚Text' von synergetischen Prozessen gesteuert wird, folgen wir der weiterführenden Annahme, daß diese Prozesse quantitativ zu beschreiben sind. Die für die einzelnen Texte erhaltenen Beschreibungsmodelle lassen sich miteinander vergleichen, was möglicherweise in einem (oder mehreren) allgemeinen Modell(en) resultiert. Auf diese Weise läßt sich letzten Endes eine quantitative Texttypologie erreichen.

2. Auch bei der Annahme einer bestimmten Definition von ‚Text' bleibt zu entscheiden, welche Text-Eigenschaften einer quantitativen Analyse unterzogen werden sollen. In der vorliegenden Studie konzentrieren wir uns exemplarisch auf die Wortlänge als eine spezifische Texteigenschaft, wobei ebenso Fragen der Definition der Einheit ‚Wort' und die Maßeinheit, in der die Wortlänge zu messen ist, darzulegen sind.

WORTLÄNGE ALS CHARAKTERISTIKUM FÜR EINE QUANTITATIVE TEXT-KLASSIFIKATION

Die Wortlänge ist bei Fragen der quantitativen Textklassifizierung als ein ausgewähltes Textcharakteristikum zu verstehen, welches für unterschiedliche Bereiche der Textwelt in Betracht gezogen werden kann. Ohne Frage gilt es im Rahmen der quantitativen Text-Kategorisierung und Texttypologie, über den ausgewählten Faktor der Wortlänge hinaus eine ganze Reihe weiterer Kenngrößen in Betracht zu ziehen, wie etwa die Satzlänge, das lexikalische Type-Token-Verhältnis u.a.m. Im Rahmen der vorliegenden Untersuchung wird allerdings die Aufmerksamkeit ausschließlich auf die Wortlänge und damit verbundene methodologische Fragestellungen eingeschränkt, damit so die Bedeutung systematischer und akribischer Vorgangsweisen nachvollziehbar demonstriert werden kann.

Bestimmung der Autorenschaft und der Wortlänge als quantitativer Parameter in Stilistik, Bestimmung der Autorschaft und der Texttypologisierung

Die Problematik der Wortlänge läßt sich aus unterschiedlichen Perspektiven diskutieren. Die Spannweite der damit verbundenen Fragestellungen reicht von der Frage der Wortlänge in einem synergetischen Regelkreis (vgl. Köhler 1986) bis hin zur Wortlänge auf Textebene, die im gegebenen Zusammenhang von unmittelbarer Bedeutung ist. Neben dieser Perspektive auf die sprachinternen Abhängigkeiten der Wortlänge von anderen – quantitativ erfaßbaren – Sprachebenen stellt die Wortlänge eine zentrale Möglichkeit dar, die Struktur von Texten statistisch-deskriptiv zu erfassen. Als damit zusammenhängende Fragestellungen sind anzusehen:

a) die zentrale Problematik der quantitativen Texttypologie und Textklassifizierung (vgl. u.a. Mistrik 1973; Alekseev 1988; Tuldava 1998)
b) die damit in engem Zusammenhang stehende Fragestellung der quantitativ erfaßbaren stilistischen Beschreibung von Texten, die sich unter dem Begriff „Stilometrie" subsumieren läßt (vgl. Fucks 1955; 1956; Doležel 1964; Martynenko 1988; u.a.)
c) der Bereich der Autor-Attribution, wo für strittige Werke/Texte die Autorschaft aufgrund der Wortlänge bestimmt werden soll (vgl. z.B. Ermolenko 1988, Marusenko 1990).

Die Wortlänge wird bei den zuletzt angesprochenen Fragestellungen allerdings nicht als eigenständiges aussagekräftiges Merkmal in Betracht gezogen. Vielmehr wird – wie einleitend bereits angedeutet – ein ganzes Set quantitativer Eigenschaften von Texten (z.B. die Satz- und Phrasenlänge bzw. allgemein die Häufigkeit von Texteinheiten wie lexikalischer Reichtum, die Textdeckung, das Type-Token Verhältnis u.ä.) in toto in empirisch-quantitative Untersuchungen einbezogen, um die Struktur von Texten quantitativ zu erfassen. Dabei ist es jedoch als problematisch anzusehen, daß durch dieses Vermischen von mehreren Eigenschaften keine Information über die Aussagekraft der Wortlänge erhalten werden kann. Nicht zuletzt deshalb ist in der vorliegenden Arbeit das Interesse ausschließlich auf die Wortlänge bzw. weitere, aus der Wortlänge abgeleitete statistische Kenngrößen ausgerichtet. Anzumerken bleibt, daß die Wortlänge in einigen wenigen Fällen (vgl. Fucks 1955; 1956; Nikonov 1978) durchaus als alleiniges Merkmal untersucht wurde – jedoch haben diese Untersuchungen hochgradig selektiven und nicht-systematischen Charakter.

Zur Frage der Untersuchungseinheit: Die Definition des Wortes und der Maßeinheiten

In jedem Fall ist für die Beschreibung von Texten auf der Basis der Wortlänge die Anwendung von a priori festzulegenden Definitionen der Untersuchungseinheiten notwendig, die einer Quantifizierung unterzogen werden.

Die erste wichtige Entscheidung betrifft die Ebene der Einheit, auf welcher die Wortlänge untersucht werden kann bzw. soll. Prinzipiell wäre zu unterscheiden zwischen den Ebenen von „Types" und „Tokens" (vgl. dazu die Studie von Alekseev 1998). Im vorliegenden Fall ist das Interesse auf die Ebene der „Tokens" gerichtet, da die Analyse – wie bei der Diskussion um die hier zugrundegelegte Definition von ‚Text' angedeutet wurde – auf

vollständige, abgeschlossene Fließ-Texte ausgerichtet ist. Insofern ist in diesem Fall das Wort als eine durch Leerstellen im Text abgegrenzte Einheit zu verstehen, d.h. daß jedes laufende Wort der untersuchten Texte auf die Wortlänge hin untersucht wird.[1] In den Texten vorkommende Abkürzungen, Zahlen, Eigennamen u.a. werden vor der hier angewandten automatisierten Wortlängen-Analyse einer Tagging-Prozedur unterzogen, sodaß z.B. Abkürzungen in vollständige grammatikalisch-morphologische Formen aufgelöst werden.

Die zweite Entscheidung betrifft die Ebene der Maßeinheit, in welcher die Wortlänge zu messen ist. Im Prinzip ist dies in unterschiedlichen Maßeinheiten möglich. So läßt sich Wortlänge durchaus in der Anzahl von

a) Buchstaben (Graphemen)
b) Phonemen (Lauten)
c) Morphemen
d) Silben

pro Wort bestimmen. Es liegt auf der Hand, daß die Bestimmung der Maßeinheit (bzw. die Entscheidung für eine bestimmte Maßeinheit) nicht ohne Auswirkung auf die nach der Bestimmung der Wortlänge durchzuführende Interpretation von Resultaten bleiben kann. Bei den angesprochenen Möglichkeiten ist davon auszugehen, daß die Messung der Wortlänge in Buchstaben (bzw. Graphemen) die leichteste und wohl deshalb auch am meisten verbreitete und am häufigsten angewandte Form der Messung darstellt. Allerdings ist insbesondere die hohe Streuung, Bi- oder gar Multimodalität der erhaltenen Häufigkeitsverteilungen, d.h. insgesamt eine sprachinterne Instabilität der erhaltenen Resultate zu beachten (vgl. dazu Grzybek & Kelih 2005a). Demgegenüber ist die Messung der Wortlänge in der Anzahl von Silben als – zumindest aus sprachinterner Sicht – weitgehend störungsfrei zu betrachten. Dies scheint auch für die Messung der Wortlänge in Morphemen zu gelten, wobei es interessanterweise zu *systematischen* statistischen Verschiebungen zwischen der Messung der Wortlänge in Silben und Morphemen kommt (vgl. Kelih 2005).

Zusammengefaßt betrachtet gelten jedenfalls die hier vorgenommenen Entscheidungen bezüglich Wortdefinition und Wahl der Silbe als Maßeinheit als Ausgangspunkt für die im Weiteren zu leistende empirisch-quantitative Textklassifizierung.

METHODEN DER QUANTITATIVEN TEXTKLASSIFIZIERUNG: THEORIE

Hat man nun sowohl die Daten der Häufigkeitsverteilungen als auch die entsprechenden Kenngrößen zu jedem einzelnen Text, lassen sich die Werte statistisch bearbeiten. Die zur Verfügung stehenden Möglichkeiten können hier nicht im einzelnen erörtert werden; dennoch sollte das Spektrum der Optionen deutlich gemacht werden, welches am einen Ende

[1] Die Diskussion um die Definition der Einheit ‚Wort' ist ohne Frage in der Linguistik als zentral anzusehen (vgl. Krámsky 1969, Wurzel 2000). Jegliche quantitativ-empirisch ausgerichtete Untersuchung impliziert jedoch das Festlegen von Kriterien, die zumindest intersubjektiv nachvollziehbar und überprüfbar sein sollten. Darüber hinaus konnte in zwei Studien gezeigt werden, daß die Wahl von unterschiedlichen Wortdefinitionen (u.a. mit Ausrichtung auf die phonetische und phonologische Ebene) gegebenenfalls eine systematische Verschiebung der statistischen Kenngrößen impliziert (vgl. Antić, Kelih & Grzybek 2005; Kelih 2005).

ausschließlich quantitative, am anderen Ende quantitativ-qualitative Vorgangsweisen beinhaltet. Damit ist nicht die im Anschluß an die Untersuchungen in jedem Fall zu stellende Frage nach der qualitativen Interpretation gemeint, sondern die Art und Weise der Einführung von qualitativen Informationen in die eigentliche quantitative Untersuchung.

1. Anwendbar ist zunächst eine ausschließlich quantitative Klassifizierung von Texten. Damit ist gemeint, daß keine qualitativen Merkmale in Form von Voraus-Informationen oder begleitenden Informationen in die Analyse eingeführt werden, sondern ausschließlich quantitative Informationen, die im gegebenen Fall aus der Wortlängenhäufigkeitsverteilung abgeleitet werden. Derartige Verfahren – die bedingt als „Tabula-Rasa-Prinzip" zu verstehen sind – laufen z.B. auf die Anwendung von sogenannten *Cluster-Analysen* hinaus. Das Prinzip von Clusteranalysen besteht darin, daß anhand von vorgegebenen Variablen (also z.B. etwa der mittleren Wortlänge oder anderer statistischer Kenngrößen) Gruppen von Fällen gebildet werden, die möglichst ähnliche Ausprägungen der Variablen (bzw. Fälle aus verschiedenen Clustern möglichst unähnliche) aufweisen. Die Anzahl der zu unterscheidenden Cluster kann in den entsprechenden Analysen vorgegeben werden; ebenso ist es aber auch möglich zu berechnen, welche Anzahl von Clustern insgesamt einer optimalen Gruppierung des gesamten Materials entspricht.
2. Zwei weitere Zugangsweisen lassen sich bedingt als a-priori/a-posteriori-Prinzip bezeichnen, insofern sie eine Synthese der Textklassifizierung auf der Grundlage von qualitativen und quantitativen Merkmalen darstellen: zum einen handelt es sich um sog. *post-hoc-Analysen*, zum anderen um sog. multivariate Diskriminanzanalysen.
   (a) In post-hoc-Analysen, die mitunter auch als a posteriori-Tests bezeichnet werden, wird im Grunde genommen die Frage beantwortet, welche Gruppen bei einem signifikante Unterschiede zwischen den untersuchten Gruppen ausweisenden Ergebnis einer Varianzanalyse für diese Signifikanz verantwortlich sind; hierbei werden solche Gruppen, zwischen denen keine signifikanten Unterschiede bestehen, zu homogenen Untergruppen zusammengefaßt. Notwendig ist also die vorherige Zuordnung der einzelnen Fälle (Texte) zu einer Gruppe, sodann die statistische Auswertung dieser Gruppe und der daran anschließende Vergleich dieser Gruppen hinsichtlich ihrer möglichen Homogenität.
   (b) Während es in den post-hoc-Tests um die Unterscheidung von homogenen Untergruppen im gesamten Datenmaterial geht, werden in *multivariaten Diskriminanzanalysen* die einzelnen Fälle (Texte) zunächst bestimmten Gruppen (wie etwa Textsorten, Funktionalstilen o.ä.) zugeordnet; in der Analyse selbst werden die einzelnen Fälle dann auf der Basis von spezifischen Prädiktorvariablen (wie etwa mittlere Wortlänge) Gruppen zugeordnet, wobei die Variablen spezifischen (in der Regel linearen) Transformationen unterzogen werden, um eine optimale Diskrimination der einzelnen Fälle zu erreichen. Auf diese Art und Weise kann getestet werden, inwiefern die qualitative a-priori-Zuordnung dem quantitativen Informationsbestand der untersuchten Texte entspricht.

EMPIRISCH-QUANTITATIVE KLASSIFIZIERUNG: TEXTBASIS UND BERECHNETE STATISTISCHE KENNGRÖßEN

Die vorliegende Untersuchung konzentriert sich auf die Analyse von insgesamt 398 slowenischen Texten. Im Hinblick auf die gewählte Methodik bei der als apriori-Klassifikation zu verstehenden Strukturierung des Text-Universums lassen sich im Prinzip verschiedene Vorgangsweisen unterscheiden, und zwar in Abhängigkeit von der jeweils zugrundegelegten Texttypologie, die in der Regel aus rein qualitativen begründeten, text-theoretischen Erwägungen abgeleitet wird. Für quantitative Untersuchungen kann jegliche qualitativ begründete Texttypologie nicht mehr und nicht weniger als *tentativen* Charakters sein. Für unsere Zwecke soll das konkret so aussehen, daß – ausgehend von der funktionalstilistischen Differenzierung der Texte (s.o.) – in einem zweiten Schritt den genannten Funktionalstilen konkrete Textsorten zugeordnet werden. Die mit diesem Vorgehen verbundene Tentativität besteht darin, daß die Zuordnung der Textsorten zu den Funktionalstilen lediglich eine apriori-Zuordnung ist und einzig und allein den Sinn hat, möglichst das gesamte Spektrum der stilistischen Differenzierung abzudecken, um sodann die Texte empirisch-quantitativ im Hinblick auf ihre stilistischen Eigenschaften zu untersuchen. Denn die mit einer rein qualitativen Typologisierung verbundene Zuordnung beinhaltet zwangsläufig eine gewisse Subjektivität. Dies haben Grzybek & Kelih (2005b) anhand einer empirischen Befragung von germanistischen Fach-Studierenden der Universität Leipzig aufgezeigt: In dieser Befragung ging es darum, mehrere Dutzend Textsorten bestimmten Funktionalstilen zuzuordnen. Interessanterweise stimmte die in *Tab. 1a* vorgenommene Zuordnung von Textsorten zu Funktionalstilen in einer Reihe von Fällen mit den Ergebnissen der Untersuchung zu 100% überein, in anderen Fällen aber konnten keine Übereinstimmungen erzielt werden. Dies verdeutlicht klar die Schwächen rein qualitativer Texttypologisierungen, denen keine quantifizierenden Textforschungen folgen.

Alltag	Wissenschaft	Administration	Journalistik	Kunst		
				Prosa	Poesie	Dramatik
1	*2*	*3*	*4*	*5*	*6*	*7*
	Abstract	Anleitung	Agenturmeldung	Autobiographie	Elegie	
Tagebucheintrag	Aufsatz	Geschäftsbrief	Auslandsbericht	Biographie	Epos	Komödie
Witz	Aufsatz_gewi	Gesetzestext	Fachartikel	Briefroman		Tragödie
	Aufsatz_nawi	Gutachten	Feuilleton	Epilog	Ode	Versdrama
	Autorreferat	Parteitagsbeschluss	Glosse	Erinnerungen	Sonett	
	Diplomarbeit	Predigt	Kolumne		Verserzählung	
	Dissertation	Schreiben		Fabel	Versroman	
	Referat	Vertrag	Kritik	Gleichnis		
	Rezension	Vortrag		Kunstmärchen		
	Tagungsbericht		Meldung	Kurzroman		
		Predigten	Sportbericht	Legende		
			Wetterbericht	Mythos		
			Zeitschriftenaufsatz	Novelle		
			Zeitungsartikel			
				Sage		
				Schwank		
				Tagebuchroman		
				Volksmärchen		

Tab. 1a: Zuordnung von Textsorten zu Funktionalstilen?

Tab. 1b veranschaulicht, wie ausgewählte Textsorten den verschiedenen Funktionalstilen zugeordnet werden können; es handelt sich um die Textsorten, die auch in der oben erwähnten Untersuchung von Grzybek & Kelih (2005b) verwendet wurden. In *Tab. 1a* sind diejenigen Textsorten unterlegt, die Gegenstand der vorliegenden Untersuchung sind, und die in *Tab. 1b* detailliert aufgeschlüsselt sind.

Die in unsere Untersuchung einfließenden Texte und ihre Struktur hinsichtlich der Funktionalstile und Textsorten ist in *Tab. 1b* anschaulich dargestellt.

Funktionalstil	Autoren	Textsorte	Anzahl
Alltag / privat	Cankar, Jurčič	Privatbrief	61
Administration	div.	Offene Briefe	29
Journalistik	div.	Leserbriefe, Kommentare	65
Prosa	Cankar	Kapitel aus Erzählungen (povest)	68
	Švigelj-Mérat / Kolšek	einzelne Briefe aus Briefroman	93
Poesie	Gregorčič	versgebundene Gedichte	40
Drama	Jančar	individuelle Akte aus Dramen	42
gesamt			398

Tab. 1b: Textbasis slowenischer Texte

Wie der *Tab. 1b* zu entnehmen ist, stellen die slowenischen Texte jeweils abgeschlossene Einzeltexte dar; weiterhin ist deutlich erkennbar, daß ein (durchaus intendierter) Fokus auf verschiedenen Briefsorten unterschiedlicher Funktionalstile liegt. Daneben ergibt sich auch die (hier nicht weiter verfolgte) Möglichkeit, den Einfluß individueller Autorschaft unabhängig von der konkreten Textsorte zu untersuchen, da von Ivan Cankar sowohl Privatbriefe als auch künstlerische Prosa vertreten sind.[2]

Berechnung statistischer Kenngrößen

Für jeden einzelnen in der *Tab. 1b* aufgeschlüsselten Text wird – unter Einhaltung der oben erwähnten Restriktionen – die Wortlänge in der Anzahl der Silben pro Wort berechnet. Die im Ergebnis erhaltenen Rohdaten (d.h. die jeweilige Anzahl *x*-silbiger Wörter in einem Text) werden in weiterer Folge in eine Wortlängenhäufigkeitsverteilung transformiert. *Abb. 2* veranschaulicht diesen Schritt, wo am Beispiel eines slowenischen journalistischen Kommentars der Anteil der *x*-silbigen Wörter bestimmt wird, sodaß sich die dargestellte Häufigkeitsverteilung ergibt.

[2] Anzumerken ist, daß die Texte vor der automatisierten Bestimmung der Wortlänge einheitlich bearbeitet werden (in Privatbriefen werden Datum, Adresse u.ä. ausgeklammert, in den Dramen werden Regieanweisungen u.ä. nicht in die Untersuchung einbezogen).

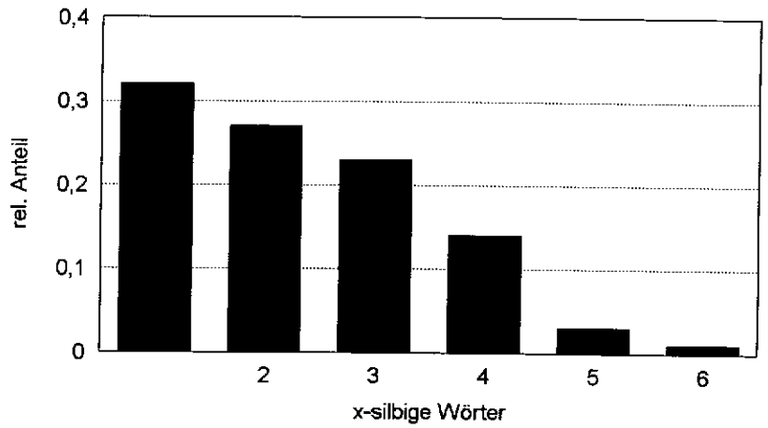


Abb. 2: Journalistischer Text (#344) mit dem relativen Anteil x-silbiger Wörter

Die Häufigkeitsverteilung der Wortlängen ist dann im weiteren der Ausgangspunkt für die Berechnung einer Vielzahl zusätzlicher statistischer Kenngrößen. Die vermutlich allgemein bekannteste Kenngröße ist sicherlich das arithmetische Mittel (d.h. im konkreten Fall die durchschnittliche Anzahl der Silben pro Wort); weitere oft verwendete Kenngrößen sind etwa die Standardabweichung bzw. deren Quadrat, die sog. Varianz (ein Maß für die durchschnittliche Abweichung vom Mittelwert). Auch Schiefe und Kurtosis, Entropie, Wiederholungsrate und andere Kenngrößen sind in der Sprach- und Textwissenschaft wiederholt zur Anwendung gekommen. Diese Kenngrößen – im Grazer Wortlängen-Projekt[3] wurde ein Set von nicht weniger als 30 solcher Kenngrößen erarbeitet – lassen sich für die entsprechenden Analysen verwenden. Nicht immer und nicht für alle Fragestellungen ist es notwendig, alle theoretisch zur Verfügung stehenden Kenngrößen zum Einsatz zu bringen. Oft reichen zwei, drei, oder maximal vier von ihnen aus, um eine gegebene Fragestellung in befriedigendem Maße zu lösen. Dabei ist man natürlich bemüht, die Anzahl der verwendeten Kenngrößen möglichst gering zu halten (damit die getroffene Aussage leichter interpretierbar wird); allerdings weiß man in der Regel nicht, zumindest nicht im vorhinein, welche Variablen im Hinblick auf den Untersuchungsgegenstand die größte Aussagekraft haben. Für die vorliegende Fragestellung einer quantitativen Texttypologisierung etwa wird sich ein Set von nicht mehr als vier Variablen als ausreichend erweisen, um eine optimale Klassifizierung zu gewährleisten – doch dazu unten mehr.

[3] http://www-gewi.uni-graz.at/quanta

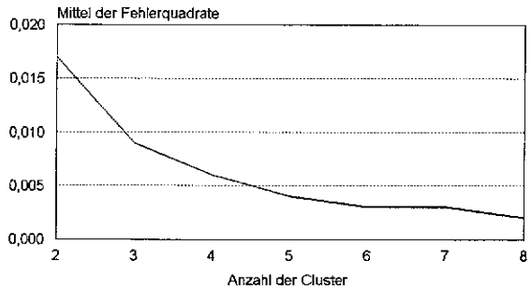
METHODEN DER QUANTITATIVEN TEXTKLASSIFIZIERUNG (EMPIRIE: 398 SLOWENISCHE TEXTE)

Cluster-Methoden

In einem ersten Schritt wird das Material einer *Cluster-Analyse* unterzogen. Dabei steht im Zusammenhang mit unserer Fragestellung das Ziel im Vordergrund, die dem gesamten Material am besten entsprechende Anzahl von Clustern zu bestimmen. Dabei wollen wir uns auf die mittlere Wortlänge der Texte als dem wesentlichsten Maß der zentralen Tendenz beschränken.

Ein gängiges Verfahren zur Bestimmung der optimalen Cluster-Anzahl ist die sog. Ellbow-Technik. Hierbei handelt es sich um ein Verfahren, das auf dem Mittel der Fehlerquadrate der Varianzanalyse für die jeweilige Clusteranzahl basiert (die bei den entsprechenden Analysen veränderlich vorgegeben werden kann). Die in *Tab. 2* enthaltenen Werte für die Clusterzahlen 3–8 sind in der *Abb. 3* graphisch dargestellt. Bei der graphischen Darstellung wird in einem *xy*-Koordinatensystem die Clusteranzahl gegen die Fehlerquadratsumme abgetragen, und an der Stelle des sich ergebenden Linienverlaufs, an der es zu

Anzahl der Cluster	Mittel der Fehlerquadrate
2	0,017
3	0,009
4	0,006
5	0,004
6	0,003
7	0,003
8	0,002



Tab. 2 / Abb. 3: Bestimmung der optimalen Cluster-Anzahl

einem überproportionalen Abfall kommt (dem Ellbow-Knick), liegt die „beste" Clusteranzahl für die gegebe Objektmenge (vgl. *Abb. 3*).

Im Vergleich dazu handelt es sich bei den sog. Two-Step-Analysen um eine explorative Prozedur zur Ermittlung von Gruppierungen innerhalb eines Datensatzes; hierbei können verschiedene Distanzmaße verwendet werden, um die (Un-)Ähnlichkeiten zwischen zwei Clustern zu berechnen. In unserem Fall führt die Verwendung des Log-Likelihood-Distanzmaßes zu demselben Ergebnis wie die Ellbow-Technik: Demnach entsprechen *drei Untergruppen* einer optimalen Gliederung des Textmaterials (vgl. *Tab. 3*).

Allein aufgrund dieses Befundes gelangt man bereits zu ersten zentralen Schlußfolgerung: Demnach deckt sich nämlich die Anzahl der berechneten Cluster nicht mit der

Zentroide			
		$\bar{x}$	
		$\bar{x}$	$\bar{s}$
Cluster	1	2,402	0,1293
	2	1,8114	0,0885
	3	2,045	0,0794
	Kombiniert	1,9889	0,2379

Tab. 3: Clusteranzahl-Analysen

traditionellen Einteilung nach sechs Funktionalstilen. Aufgrund der offensichtlichen Inadäquatheit der Funktionalstile liegt es deshalb nahe, im Hinblick auf die qualitative a-priori-Zuordnung auf eine tiefere Ebene zurückzugehen und zu untersuchen, ob nicht gegebe-nenfalls die Ebene der Textsorten, anstelle die der Funktionalstile, zutreffendere Aussagen über die Strukturierung des Textuniversums ermöglicht.

Post-hoc-Verfahren der Klassifizierung

Wie oben bereits erwähnt, kann mit Hilfe sogenannter *post-hoc-Analysen*, die im gegebenen Fall allein auf dem arithmetischen Mittel beruhen, die Frage beantwortet werden, welche Textsorten sich zu „homogenen Untergruppen" zusammenfassen lassen. Bezogen auf unser Korpus der 398 slowenischen Texte bilden sich statistisch - wie aus *Tab. 4* ersichtlich - *fünf homogene Untergruppen*.

Funktionalstil	Untergruppen für $\alpha = 0{,}05$				
	1	2	3	4	5
Gedichte	1,7127				
Erzählungen (povest)		1,8258			
Privatbriefe		1,8798			
Drama		1,8973			
Briefromane			2,0026		
Leserbriefe				2,2622	
Kommentare				2,2883	
Offene Briefe					2,4268

Tab. 4: Resultate der post-hoc-Analysen

Als erstes gilt es festzuhalten, daß eine Reihe von Textsorten aufgrund einer ähnlichen Mittelwertstruktur gemeinsame homogene Untergruppen bildet. Dabei ist eine Beobachtung von besonderem Interesse, nämlich daß die vier untersuchten Brieftypen in vier unterschiedliche Untergruppen fallen. Folglich ist davon auszugehen, daß der Typus ‚Brief' nicht eine undifferenzierte einheitliche Textsorte repräsentiert, sondern in sich ein breites Spektrum möglicher Funktionalstile abdeckt. Dies ist deshalb von Bedeutung, als der ‚Brief' in zahlreichen quantitativen Untersuchungen quasi als sprachlicher Prototyp gehandelt wird, da er zum einen als Ergebnis eines homogenen Prozesses der Textgenerierung, zum anderen als ideale Mischung zwischen sprachlicher und mündlicher Kommunikation angesehen wird.

Von Bedeutung ist weiterhin auch die Beobachtung, daß insgesamt fünf homogene Untergruppen unterschieden werden - dies deckt sich interessanterweise nicht mit dem Ergebnis der Clusteranalyse, der zufolge die Zusammenfassung aller Texte in insgesamt drei Gruppen eine optimale Kategorisierung wäre. Das kann in der Interpretation nur bedeuten, daß sich einige der Textsorten nicht (bzw. nicht allein) aufgrund der Wortlänge differenzieren lassen, sondern in ein übergeordnetes Cluster (eine abstraktere Textkategorie) eingehen. Theoretisch wäre es möglich, daß sich hier eine Zuordnung der Textsorten zu den Funktionalstilen ergibt - möglich wäre es aber auch, daß die Obergruppen von ganz anderer Qualität sind und in einer neuen Art von Typologie resultieren, etwa einer spezifischen Diskurs-Typologie.

Aufschluß darüber geben können nur *multivariate Diskriminanzanalysen*: Während post-hoc-Analysen auf die Unterscheidung homogener Untergruppen abzielen, geht es in Diskriminanzanalysen darum, die einzelnen Fälle (d.h. die einzelnen Texte) auf der Basis von spezifischen Prädiktorvariablen bestimmten Gruppen zuzuordnen; als Prädiktorvariablen fungieren in unserem Fall die Wortlänge bzw. aus der Wortlängenhäufigkeitsverteilung abgeleitete statistische Kenngrößen. Im Gegensatz zu den post-hoc-Analysen liegt den multivariaten Diskriminanzanalysen nicht nur jeweils eine Variable zugrunde, sondern ein Variablenbündel. Dabei werden die Variablen so transformiert, daß am Ende eine optimale Diskrimination der einzelnen Fälle herauskommt: Im Ergebnis läßt sich dann sagen, wie viele der Fälle (bzw. wieviel Prozent der Fälle) aufgrund der verwendeten Kenngrößen den a priori zugeordneten Kategorien „richtig" zugeordnet werden, was als Indiz für die Güte der gewählten Prädiktorvariable(n) gewertet werden kann.

Diskriminanzanalysen zur Textklassifizierung

Beim qualitativ-quantitativen Verfahren der Diskriminanzanalyse läßt sich also einerseits die Wahl der Prädiktorvariablen in Abhängigkeit von deren Aussagekraft ändern; andererseits lassen sich die Analysen natürlich auch auf der Basis unterschiedlicher qualitativer a priori-Zuordnungen (etwa zu Textsorten oder Funktionalstilen) durchführen.

In unserem Fall ist in einem ersten Schritt das Verfahren der multivariaten Diskriminanzanalyse auf der Ebene der unterschiedenen Funktionalstile anzuwenden. In diesem Fall wird davon ausgegangen, daß den untersuchten Texten insgesamt vier Funktionalstile zugrundeliegen (die Textsorten aus dem Bereich des Dramas, des Verses und der Prosa werden zum Funktionalstil „Kunst" zusammengefaßt). Wie der *Tab. 5* zu entnehmen ist, können die Texte unter dieser Voraussetzung aufgrund von zwei Variablen (m_1 und o_5) mit einem Prozentsatz von 74.40% korrekt zugeordnet werden.[4] Wie aus der graphischen Darstellung ersichtlich ist, bilden sich zusammengefaßt zwei große Gruppen von Texten, die sich zu weiten Teilen überschneiden: Die Texte aus dem Bereich des „Alltags" verstehen sich in dieser Klassifikation als „Künstlerischer Stil". Erklärbar ist dies unter Umständen aufgrund der Tatsache, daß sich Privat-Briefe nicht von anderen künstlerischen Texten unterscheiden.

[4] Die Variable m_1 ist das erste (dem Mittelwert entsprechende) Zentralmoment der Verteilung, o_5 ist der Quotient aus dem 2. und 1. Zentralmoment (der auch als Ord'sches O bezeichnet wird).

Funktionalstil	Vorhergesagte Gruppenzugehörigkeit				
	Alltag	*administrativ*	*journalistisch*	*Kunst*	*gesamt*
Alltag	1	0	2	58	61
administrativ	0	18	8	3	29
journalistisch	0	9	42	14	65
Kunst	2	0	6	235	243

Tab. 5: Diskriminanzanalyse: Vier Funktionalstile (n = 398)

Ähnliche Überlappungen ergeben sich auf der Ebene von administrativen und journalistischen Texten, die aufgrund der Wortlänge nicht zufriedenstellend zu trennen sind.

Im zweiten Schritt werden nunmehr die Texte nicht mehr nach den Funktionalstilen, sondern nach den Textsorten vorab-klassifiziert und einer Diskriminanzanalyse unterzogen. Da wir es im gegebenen Fall mit acht verschiedenen Textsorten zu tun haben, muß dies der erste Schritt sein. Im Ergebnis stellt sich heraus, daß selbst auf der Basis von vier Prädiktorvariablen – nämlich dem arithmetischen Mittel (m_1), der Varianz (m_2), dem relativen Anteil einsilbiger Wörter (p_1), sowie dem als Quotient aus der Standardabweichung s und dem Mittelwert m_1 berechneten Variationskoeffizienten v – eine nur zu 56.30% korrekte Zuordnung der Texten möglich ist vgl. *Tab. 6* – die Berücksichtigung weiterer Variablen führt zu keiner substantiellen Verbesserung des Resultats. Dieser Sachverhalt ist in *Abb. 5* dargestellt.

Bereits an dieser Stelle sind einige Tendenzen deutlich zu erkennen: So nehmen alle

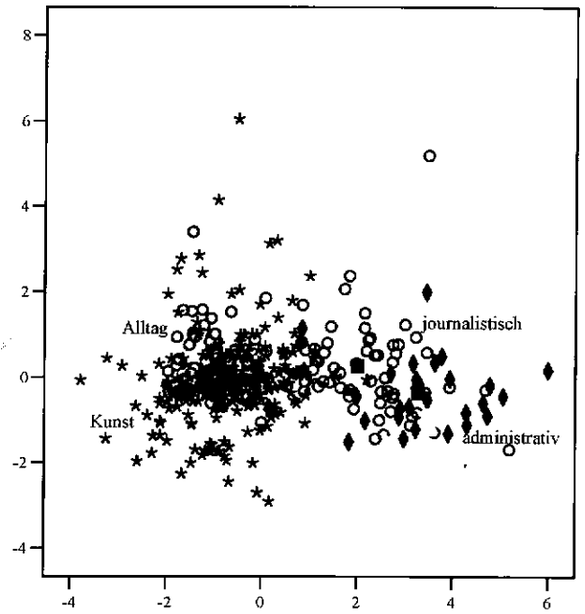


Abb. 4: Diskrimination von vier Funktionalstilen (n = 398; 74.40%; Variablen: m_1 und o_s)

Textsorte	Vorhergesagte Gruppenzugehörigkeit								
	PB	*OB*	*LB*	*KO*	*BR*	*ER*	*GE*	*DR*	*gesamt*
PB	22	0	0	0	19	20	0	0	61
OB	0	21	0	5	3	0	0	0	29
LB	1	6	3	10	10	0	0	0	30
KO	0	7	1	16	11	0	0	0	35
BR	3	0	1	2	78	8	1	0	93
ER	11	0	0	0	8	49	0	0	68
GE	0	0	0	0	0	6	34	0	40
DR	5	0	0	0	19	16	1	1	42

Tab. 6: Diskriminanzanalyse: Vier Funktionalstile (n = 398)

Textsorten jeweils ein mehr oder weniger klar abgegrenztes Feld ein: Die versgebundenen Gedichte nehmen einen deutlich abgetrennten Bereich ein; ähnliches gilt auch für Leserbriefe, offene Briefe und Kommentare. Drama, Erzählungen, Privat-Briefe sowie die Briefe aus dem Briefroman nehmen im Gegensatz dazu ein etwas weiteres, nicht ganz so deutlich abgegrenztes Feld ein, wobei sich die einzelnen Texte dieser Textsorten nicht klar voneinander abgrenzen. Dieses Ergebnis führt zu der Überlegung, das Textmaterial einzugrenzen, um auf diese Art und Weise bestimmten Grundmustern des Textuniversums auf die Spur zu kommen. Als eine erste Einschränkung bietet sich eine reduzierte Betrachtung

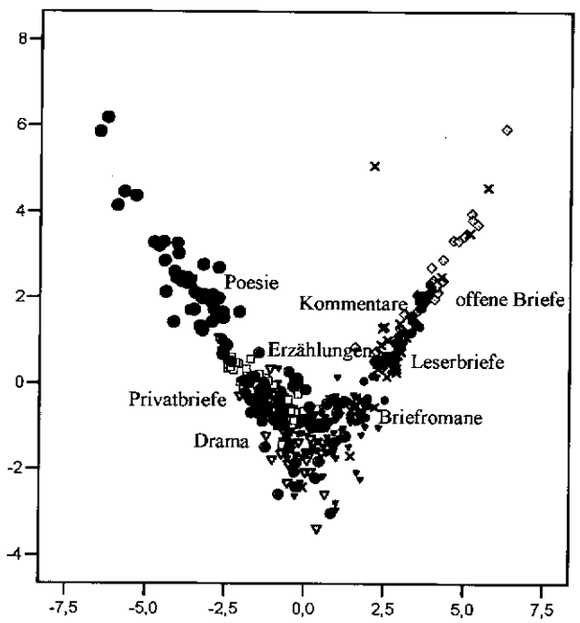


Abb. 5: Diskriminanzanalyse: Acht Textsorten (n = 298; 56.30%; Variablen: m_1, m_2, v, p_1)

der vier verschiedenen Briefsorten - Privatbriefe (PB), offene Briefe (OB), Leserbriefe (LB) und Briefe aus dem Briefroman (BR) an -, die ja aufgrund der vorangehenden post-hoc-Analysen in vier verschiedene Untergruppen getrennt wurden.

Abermals bekommt man mit der multivariaten Diskriminanzanalyse mit zwei Variablen - in diesem Fall dem Mittelwert (m_1) und v - eine korrekte Zuordnung von 70.40% der Texte. Die Ergebnisse sind in *Tab. 7* dargestellt.

Aus *Tab. 7* ist ersichtlich, daß die Privatbriefe (PB) und die Briefe aus dem Briefroman (BR) die Tendenz haben, ineinander überzugehen und eine gemeinsame Gruppe zu bilden. Im Vergleich dazu bilden die offenen Briefe (OB) einen relativ eigenständigen Bereich, wo-

	Vorhergesagte Gruppenzugehörigkeit				
Brieftypen	*PB*	*OB*	*LB*	*BR*	*gesamt*
PB	37	0	2	22	61
OB	0	22	3	4	29
LB	1	9	10	10	30
BR	10	0	3	80	93

Tab. 7: Diskriminanzanalyse: Vier Brieftypen (n = 213)

hingegen die Leserbriefe (LB) sich zu drei fast gleich großen Anteilen auch auf den Bereich der offenen Briefe bzw. den des Briefromans verteilen. Formiert man deshalb eine gemeinsame, aus Privatbriefen und Briefromanen bestehende Gruppe und stellt diese den Leserbriefen und den offenen Briefen gegenüber, so erhält man eine beachtliche Trennung von immerhin 86.90%. Die entsprechenden Ergebnisse sind in *Tab. 8* zusammengefaßt.

Zu sehen ist, daß die kombinierte Gruppe 1 (Privatbriefe und Leserbriefe) zu ca. 98% korrekt unterschieden wird. Dies ist als starkes Argument dafür zu werten, daß wir es hier mit einer allgemeinen Gruppe von privaten Briefen zu tun haben, wobei es vernachlässig-

	Vorhergesagte Gruppengröße			
Gruppe	*1*	*2*	*3*	*gesamt*
1	151	0	3	154
2	2	20	6	28
3	12	5	14	31
	1 = {PB,BR}	2 = OB	3 = LB	

Tab. 8: Diskriminanzanalyse: Drei Brieftypen (n = 213)

	Vorhergesagte Gruppengröße		
Gruppe	*1*	*2*	*gesamt*
1	152	3	154
2	14	45	59
	1 = {PB,BR}	2 = OB, LB	

Tab. 9: Diskriminanzanalyse: Zwei Brieftypen (n = 213)

bar ist, ob diese literarischer Provenienz sind oder nicht. In Anbetracht dieses Befunds ist daher anzuzweifeln, ob aus quantitativer Sicht (bzw. auf der Basis der Wortlänge) eine Trennung von literarischen Briefen und Briefen aus Briefromanen möglich ist; anzunehmen ist stattdessen, daß die literarischen Briefe den Stil von Privatbriefen imitieren bzw. reproduzieren; diese Beobachtung läuft jedoch in letzter Konsequenz darauf hinaus, daß gegebenenfalls die

Existenz von künstlerischer Literatur als eigener Funktionalstil (bzw. als eigene Funktionalstile) grundsätzlich in Frage zu stellen sein wird.

Ungeachtet dieser offenen Frage liegt es nunmehr nahe, auch die Leserbriefe und die offenen Briefe in einer gemeinsamen Gruppe zusammenzufassen. Die Diskriminanzanalyse führt in diesem Fall (Gruppe 1: Leserbriefe und offene Briefe, Gruppe 2: Privatbriefe und Briefe aus dem Briefroman) auf der Basis von zwei Variablen - der mittleren Wortlänge m_1 und dem Anteil zweisilbiger Wörter p_2 - zu einer ca. 92% korrekten Trennung (vgl. *Tab. 9*) dieser Gruppen, die als überaus zufriedenstellend anzusehen ist.

AUF DEM WEG ZU EINER NEUEN TEXTTYPOLOGIE

Es liegt natürlich auf der Hand, daß eine Reduktion des Materials und eine Reduktion der untersuchten Kategorien zu einem insgesamt besseren Resultat führen. Ungeachtet dessen stellt sich auf der Grundlage der obigen Beobachtungen die Frage, ob die vorgenommene Zweiteilung - auf der einen Seite private Briefe (PB, BR), auf der anderen Seite öffentliche Briefe (LB, OB) - einen Spezialfall von zwei allgemeineren Kategorien darstellt: nämlich eines Privat- bzw. Alltagsstils einerseits und eines öffentlichen bzw. offiziellen Stils andererseits. Wenn diese Hypothese den Tatsachen entspricht, dann sollte die entsprechende Wiedereinführung der vorübergehend aus der Untersuchung ausgeschlossenen Textsorten zu einer deutlichen Verbesserung der obigen Resultate führen. Letztendlich gilt es natürlich, alle Texte wieder in ein allgemeines Modell einzuführen, dessen Gruppierung freilich kaum der Ausgangsstruktur entsprechen wird.

	Vorhergesagte Gruppengröße		
Gruppe	*1*	*2*	*gesamt*
1	148	6	154
2	16	78	94
	1 = {PB, BR}	2 = {OB, LB, KO}	

Tab. 10: Diskriminanzanalyse: Fünf Textsorten in zwei Kategorien: (öffentlich/offiziell vs. Privat- bzw. Alltagsstil, n = 248)

Tatsächlich läßt sich - wie aus *Tab. 10* ersichtlich - zeigen, daß sich auch bei Wiedereinführung der journalistischen Kommentare (KO) der Anteil von korrekt getrennten Texten nicht wesentlich verschlechtert. Von den nunmehr 248 Texten werden weiterhin 91.10% korrekt klassifiziert. Es verstärkt sich damit die Annahme, daß die Trennung von öffentlichen/offiziellen vs. Privat- bzw. Alltagstil von hoher Relevanz ist.

Analog dazu können in einem weiteren Schritt auch die Drama-Texte wie-derum in die Analyse einbezogen werden. Damit werden bereits wieder 290 Texte in die Analyse einbezogen, wobei weiterhin die Teilung in zwei große Gruppen bestehen bleibt. Zu erwarten ist hierbei, daß die Dramen-Texte - in Analogie zu den Briefen aus dem Briefroman - als literarisches Pendant von Alltagsdialogen aufzufassen (und gegebenenfalls von anderen stark dialogisch strukturieren Texten nicht aufgrund ihrer Wortlänge zu unterscheiden) sind. Im Ergebnis stellt sich eine korrekte Trennung von 92.40% der Texte heraus (vgl. *Tab. 11*).

Insgesamt zeigt sich also, daß die Wiedereinführung von zwei zuvor eliminierten Textsorten, die keinerlei Bezug zum Typus Brief haben, keinesfalls eine Verschlechterung der

Gruppe	Vorhergesagte Gruppengröße		
	1	*2*	*gesamt*
1	190	6	196
2	16	78	94
	1 = {PB, BR; DR}	2 = {OB, LB, KO}	

Tab. 11: Diskriminanzanalyse: Sechs Textsorten in zwei Kategorien: (öffentlich/ offiziell vs. Privat- bzw. Alltagsstil, n = 290)

vorangehenden Ergebnisse bewirken. An dieser Stelle sollen deshalb nun ebenfalls die Gedichte (GE) wieder in die Analyse eingeführt werden, die ja bei den post-hoc Analysen eine separate homogene Untergruppe bildeten. In der Tat führt auch dieser Schritt, der die Textbasis auf eine Anzahl von immerhin 330 Texten erhöht, zu keiner Verschlechterung des Ergebnisses. Dabei sind unter dieser Voraussetzung die folgenden drei Variablen für eine maximale Trennung verantwortlich: der Mittelwert (m_1), der relative Anteil von zweisilbigen Wörtern (p_2), sowie der Variationskoeffizient v. Die detaillierten Ergebnisse finden sich in *Tab. 12*.

Gruppe	Vorhergesagte Gruppengröße			
	1	*2*	*3*	*gesamt*
1	191	3	2	196
2	19	75	0	94
3	5	0	35	40
	1 = {PB, BR; DR}	2 = {OB, LB, KO}	3 = GE	

Tab. 12: Diskriminanzanalyse: Sieben Textsorten in drei Kategorien: (öffentlich/offiziell, Privat- und Alltagsstil, Gedichte, n = 330)

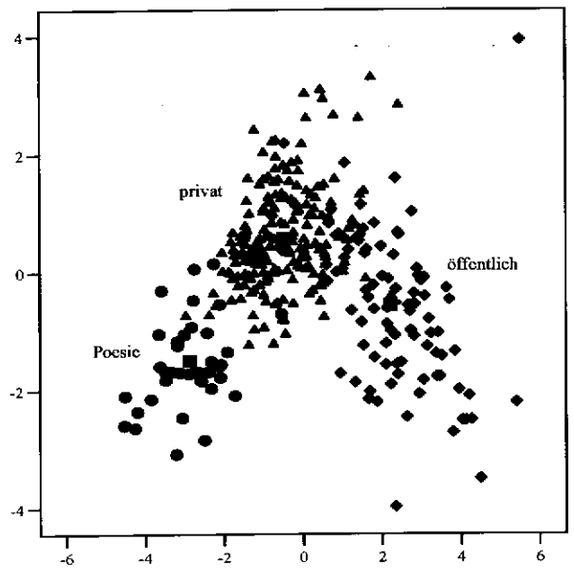


Abb. 6: Diskriminanzanalyse: Drei Diskurstypen (n = 330; 91.20%; Variablen: m_1, p_2, v)

Gruppe	Vorhergesagte Gruppengröße				
	1	*2*	*3*	*4*	*gesamt*
1	183	3	9	1	196
2	19	75	0	0	94
3	42	0	26	0	68
4	1	0	5	34	40
	1= PB, BR;DR	2 = OB, LB, KO	3 = EZ	4 = GE	

Tab. 13: Diskriminanzanalyse: Acht Textsorten in vier Kategorien: (öffentlich/ offiziell, Privat- bzw. Alltagsstil, Gedichte, literarische Erzählungen n = 398)

Die Resultate sind in *Abb. 6* anschaulich dargestellt. Es ist klar ersichtlich, daß die poetischen Texte (GE) einerseits und die öffentlichen Texte andererseits zwei Extrempunkte des gesamten Textspektrums einnehmen.

Weiterhin ist zu sehen, daß es zu keinen nennenswerten Fehl-Klassifikationen kommt. Damit stellt sich die berechtigte Frage, ob die sich ergebende Typologie bereits das Endresultat der quantitativen Textklassifikation darstellt. Zu beachten ist jedoch, daß erst sieben von acht Textsorten in die Analyse einbezogen wurden. In einem letzten Schritt sollen daher auch die Texte der Textsorte ‚Erzählungen' (EZ) wieder in die Analyse eingeführt werden. Wie aus *Tab. 13* hervorgeht, sinkt der Anteil von korrekt klassifizierten Texten durch diese Wiedereinführung der achten Textsorte, die als eigenständige Gruppe definiert wird, auf 79.80%.

Bei näherer Betrachtung der Resultate zeigt sich der Grund für die Verschlechterung des Gesamtergebnisses: Es stellt sich nämlich heraus, daß die Fehlklassifizierungen vor allem die literarischen Erzählungen und die privaten Briefe betreffen. Im Hinblick auf eine Interpretation dieses Befundes liegt es nahe, daß hier die Frage der Autorschaft ins Spiel kommt: schließlich stammen sowohl die privaten Briefe als auch die literarischen Erzählungen von ein und demselben Autor, nämlich von Ivan Cankar. Diese Interpretation würde jedoch voraussetzen, daß Autorschaft einen wichtigen Einfluß auf die Wortlänge ausübt und deswegen ein wichtiger Faktor der Textklassifizierung auf der Basis von Wortlängenanalysen ist. Daß dies jedoch nicht der Fall ist, konnte andernorts in diesbezüglichen Detailuntersuchungen gezeigt werden (Grzybek et al. 2005; Kelih et al. 2005).

Stattdessen bietet sich eine alternative Interpretation an, der folgende Überlegung zugrunde liegt: Bei den analysierten slowenischen Erzählungen handelt es sich um Texte, die zum einen relativ viele Dialogsequenzen beinhalten und die zum anderen in den narrativen bzw. deskriptiven Passagen stark mündliche Rede (im Sinne der *skaz*-Theorie des Russischen Formalismus) imitieren. Dies ließe sich dahingehend interpretieren, daß diese Texte eher einem mündlichen Sprachgebrauch entsprechen (was allerdings mitnichten bei literarischen Texten bzw. Erzählungen grundsätzlich der Fall sein muß). Jedenfalls würde dies durchaus eine Zuordnung dieser Texte zur Gruppe des Privat- bzw. Alltagsstils rechtfertigen. Im Ergebnis zeigt sich jedenfalls eine zu 92.70% korrekt durchgeführte Trennung in drei große homogene Gruppen (vgl. die Ergebnisse in *Tab. 14*).

Das Endergebnis der vorliegenden Untersuchung ist in *Abb. 7* dargestellt. Es kann aufgrund der Untersuchungen zusammenfassend festgehalten werden, daß sich in letzter Instanz das vorliegende Material am besten in drei unterschiedliche Diskurstypen eintei-

Gruppe	Vorhergesagte Gruppengröße			
	1	*2*	*3*	*gesamt*
1	260	3	1	264
2	19	75	0	94
3	6	0	34	40
	1 = PB, BR; DR, EZ	2= OB, LB, KO	3 = GE	

Tab. 14: Diskriminanz-Analyse: Acht Textsorten in drei Kategorien: Öffentlich/offiziell, Privat- bzw. Alltagsstil, Gedichte, *n* = 398)

len läßt. Diese Dreiteilung entspricht dem Eingangsbefund der Clusteranalysen, deckt sich allerdings nicht mit der traditionellen Einteilung nach Funktionalstilen.

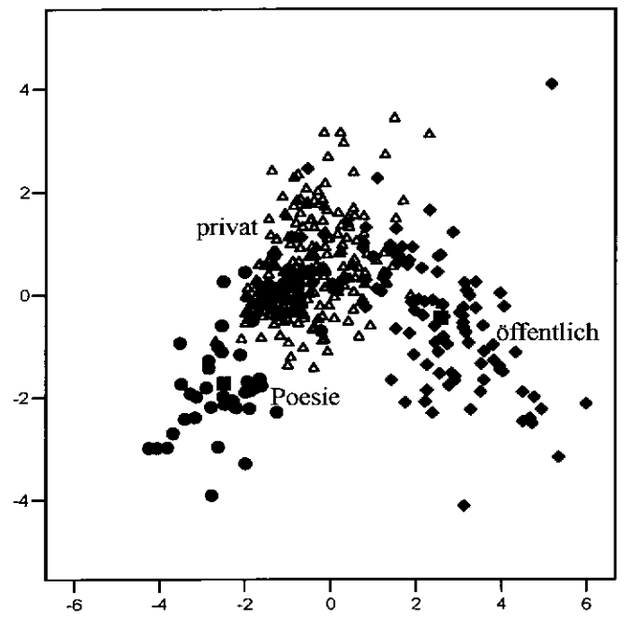


Abb. 7: Diskriminanzanalysen: Drei Diskurstypen (*n* = 398; 92.70%; Variablen: m_1, p_2, *v*)

An dieser Stelle ist ein vorläufiges Fazit unserer Untersuchungen zu ziehen: 398 slowenische Texte konnten mit Hilfe multivariater Diskriminanzanalysen zu entsprechenden Diskurstypen mit beachtlichen 92.70% zugeordnet werden, womit sich herausstellt, daß die Wortlänge ein überaus geeignetes Instrument zur Klassifikation von Texten ist.

Die sich aufgrund dieses überzeugenden Befunds zwangsläufig stellende Frage lautet, ob daraus ein allgemein gültiges Prinzip der Textstrukturierung abzuleiten ist. Daß dies nur bedingt gesagt werden kann, soll die folgende abschließende Überlegung zeigen, mit der nicht zuletzt die prinzipielle Offenheit des diskutierten Verfahrens demonstriert werden kann, indem die Einführung einer weiteren Textsorte vorgenommen wird.

Zu diesem Zweck sollen zu den bislang betrachteten 398 Texten 30 slowenische Kochrezepte hinzukommen. Betrachtet man diese als einer eigenständigen Textkategorie zugehörig - die in etwa einem Bereich wie Fachsprache oder Gebrauchs-Stil entspricht -, so bilden sich in der Tat nicht mehr nur drei, sondern vier zentrale Gruppen heraus (vgl. *Abb. 8*), die nach wie vor in der Höhe von 91.80 % zu trennen sind.

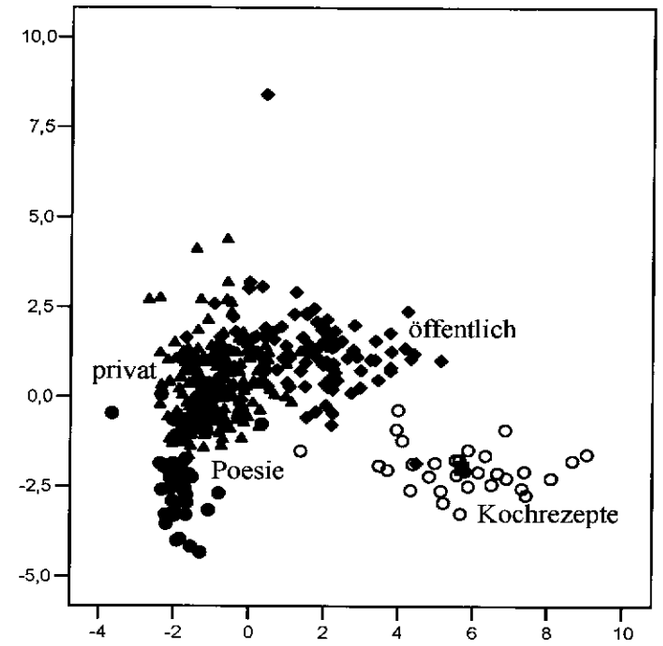


Abb. 8: Diskriminanzanalyse: Vier Diskurstypen (*n* = 429; 91.80%; Variablen: m_2, p_4, *v*)

Das bedeutet, daß selbst das Hinzufügen einer weiteren Textsorte (unter Verwendung der Variablen m_2, p_4, *v)* nichts an dem empirischen Faktum ändert, daß jeweils über 90% der Texte korrekt den verschiedenen definierten Diskurstypen zugeordnet werden können. Inwiefern dieser Befund bei der Analyse von weiteren Texten aufrechtzuerhalten ist, oder inwiefern sich die hieraus ergebende Klassifikation auch bei Texten in anderen Sprachen als von Relevanz erweist, oder gar mit der sich hier ergebenden Typologie deckt, können nur weitere systematische Untersuchungen zeigen - das ist das Schicksal, aber auch der Vorteil von empirischen Methoden...

ZUSAMMENFASSUNG

Bei der vorliegenden Untersuchung handelt es sich um einen Versuch, eine Menge von Texten - konkret werden 398 slowenische Texte analysiert - durch die Quantifizierung von Textstrukturen zu kategorisieren. Ausgehend von einem Vergleich mit dem Universum

wird die Textmenge als Text-Universum bezeichnet. Da für die beabsichtigte Strukturierung des Universums eine Identifikation der Objekte notwendige Voraussetzung ist, beinhaltet der erste Schritt eine theoriebezogene Diskussion sowohl des Wort als auch des Textbegriffs. Die weiterführende Klassifikation dieser Objekte resultiert in der Identifikation und Beschreibung hierarchisch geordneter Subsysteme: Unter Fortführung der astronomischen Analogie gilt es als nächstes, innerhalb des Text-Universums möglicherweise existierende (Text-)Galaxien zu identifizieren, welche Attraktoren für die individuellen (Text-)Objekte darstellen, wobei innerhalb solcher Galaxien spezifische Sub-Systeme niederer Ordnung zu identifizieren sind, die in Analogie zu Stern- oder Sonnensystemen zu sehen sind. Konkret auf die Texte bezogen, ergibt sich somit die Möglichkeit einer Strukturierung des Textuniversums nach Kategorien verschiedener hiearchischer Ebenen:

1. Funktionalstile
2. Textsorten
3. Einzeltexte

Bei der operationalen Umsetzung des aufgezeigten Ziels kommen unterschiedliche Methoden der quantitativen Analyse zur Anwendung, insbesondere Cluster-Analysen, post-hoc-Analysen, multivariate Diskriminanzanalysen. Als Ergebnis dieser Untersuchungen stellt sich heraus, daß eine Gruppierung der Texte nach Funktionalstilen den Ergebnissen der quantitativen Untersuchungen nicht standhält. Stattdessen ergibt sich die Notwendigkeit und Möglichkeit, das gesamte Text-Universum in drei bzw. vier Diskurstypen zu untergliedern, deren Relevanz und Tragfähigkeit sowohl im Hinblick auf weitere Texte als auch auf andere Sprachen zu prüfen sein wird.

LITERATUR

Adamzik, K.
1995 *Textsorten - Texttypologie. Eine kommentierte Bibliographie.* Münster: Nodus.

Alekseev, P.M.
1988 *Kvantitativnaja lingvistika teksta.* Leningrad: ILU.
1998 „Graphemic and syllabic length of words in text and vocabulary". *Journal of Quantitative Linguistics* 5(1-2): 5-12.

Antić, G., E. Kelih & P. Grzybek
2004 „Zero-syllable words in determining word length". In: Grzybek, P. (ed.). *Contributions to the Science of Language. Word Length Studies and Related Issues.* Dordrecht, NL: Kluwer [in print].

Blühdorn, H.
1990 „Korpuslinguistische Befunde als Ausgangspunkt für eine modifizierte Funktionalstilistik - Anregungen zu einer Neuaufnahme der Diskussion". *Linguistische Berichte* 127: 217-231.

Doležel, L.
1964 „Verojatnostnyj podchod k teorii chudožestvennogo stilja". *Voprosy jazykoznanija* 1: 19-29.

Ermolenko, G.V.
1988 *Anonimnye proizvedenija i ich avtory.* Minsk: IMU.

Fucks, W.
1955 *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen* (=Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen 34a). Köln-Opladen.
1956 *Zur Deutung einfachster mathematischer Sprachcharakteristiken* (=Forschungsberichte des Wirtschafts- und Verkehrsministerium Nordrhein-Westfalen 344). Köln-Opladen.

Grzybek, P. & E. Kelih
2005a „Warum berechnen wir die Wortlänge eigentlich (nicht) in Graphemen?" In: dies. (Hg.). *Wörter - Längen - Häufigkeiten* [in Arbeit].
2005b „Textforschung: Empirisch!" In: Banke, J., B. Dumont & A. Schröter (Hg.). *Textsortenforschungen.* Leipzig, 13-14.

Grzybek, P. & E. Stadlober
2003 „Zur Prosa Karel Čapeks - Einige quantitative Bemerkungen". In: Kempgen, S., U. Schweier & B. Tilmann (Hg.), *Rusistika - Slavistika - Lingvistika. Festschrift für Werner Lehfeldt zum 60. Geburtstag.* München: Sagner, 474-488.

Grzybek, P., E. Stadlober, E. Kelih & G. Antić
2005 „Quantitative text typology: the impact of word length". In: Weihs, C. & W. Goul (Hg.). *Classification - The Ubiquitous Challenge.* Heidelberg-Berlin: Springer [in print].

Havránek, B.
1976 (1942) „Die funktionale Schichtung der Literatursprache". In: *Grundlagen der Sprachkultur. Beiträge der Prager Linguistik zur Sprachtheorie und Sprachpflege.* Teil 1. Berlin (DDR), 150-161.

Kelih, E., G. Antić, P. Grzybek & E. Stadlober
2005 „Classification of author and(or) text?" In: Weihs, C. & W. Goul (Hg.). *Classification - The Ubiquitous Challenge.* Heidelberg, Springer, 498-505.

Kelih, E.
2005 „Einheiten der Berechnung von Wortlängen: Morphem vs. Silbe". In: Grzybek, P. & E. Kelih (Hg.). *Wörter - Längen - Häufigkeiten* [in Arbeit].

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

KRÁMSKY, J.
1969 *The Word as a Linguistic Unit*. The Hague: Mouton.

MARUSENKO, M.A.
1990 *Atribucija anonimnych i psevdoanonimnych literaturnych proizvedenij metodami raspoznavanija obrazov*. Leningrad: ILU.

MARTYNENKO, G.J.
1988 *Osnovy stilemetrii*. Leningrad: ILU.

MISTRÍK, J.
1973 *Exakte Typologie von Texten* (=Arbeiten und Texte zur Slavistik 3). München: Sagner.

NIKONOV, V.A.
1978 „Dlina slova". *Voprosy jazykoznanija* 6: 104–111.

RIESEL, E.G. & E.I. SCHENDELS (RIZEL', E.G. & E.I. ŠENDEL'S)
1975 *Deutsche Stilistik*. Moskau: Verlag „Hochschule".

SCHARNHORST, J.
1981 „Zum Wesen des Begriffs Funktionalstil". *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 34(3): 305–314.

STEUBE, A.
1974 „Funktionalstilistische Differenzierung der Sprache". *Linguistische Arbeitsberichte* 10: 114–119.

TULDAVA, J.
1998 *Probleme und Methoden der quantitativ-systemischen Lexikologie* (=Quantitative Linguistics 59). Trier: wvt.

WURZEL, W.U.
2000 „Was ist ein Wort?" In: Thieroff, R. et al. (Hg.). *Deutsche Grammatik in Theorie und Praxis*. Tübingen: Niemeyer: 29–42.

LEXICAL RICHNESS IN SLAVIC TEXTS

JANA KUSENDOVÁ

Summary: *Different statistical methods, which characterize the text by exact analysis, are applied to analyze the lexical richness of text. Some of these methods are demonstrated for poems by M. Válek and a novel of Peter Holka.*

Zusammenfassung: *Gezeigt werden verschiedene statistische Methoden, die für die Analyse des lexikalischen Reichtums in slawischen Texten anwendbar sind. Die Methoden werden an Gedichten von M. Válek und einem Roman von Peter Holka demonstriert.*

A multitude of methods have been invented to provide characterizations of texts. The methods of mathematics and statistics, which characterize the text by the exact analysis, belong to the modern ones. For mathematicians, the common linguistic practice of developing linguistic terminology is unsatisfactory. They try to elaborate on this terminology with mathematical methods. Indeed, it its hardly possible to build up scientific theories of empirical disciplines without mathematics.

The research in the field of vocabulary richness of the text dates back to 1944 and is connected with the name of the mathematician E.G. Yule, who specialized in statistics. By now, this field has developed into a separate branch, mainly thanks to efforts of French philologists.

It is not easy to find a definition for the term "vocabulary richness of a text". This term is easily defined in either statistics or philology. We may consider vocabulary richness of a text as the dependence of the number of distinct words on the number of all words in a text.

To work with a given text it is important to define the criteria which we are going to follow. An exact and uniform definition of criteria is very important. When we change a criterion, the result would be changed as well. It is possible that this changing of criteria would bring better results. But then we cannot compare results because different criteria would bring us different results.

Before we define the criteria, we should answer the following question: What does it mean when we talk about "a single word". For us, a single word consists of a set covering all of its forms. We sometimes encounter problems with the term lexeme, e.g. *písací stroj* – there are two words but only one lexeme. In other languages such words are made by composition words, e.g. Ger. *Schreibmaschine*, Engl. *typewriter*, or it can be made analytically, e.g. Ital. *macchina da scrivere*. Jozef Mistrík distinguished two different terms: word configuration and grammatical configuration. In the configuration *bol by som býval*, there

TEXT & REALITY
TEXT & WIRKLICHKEIT

Edited by Jeff Bernard, Jurij Fikfak, Peter Grzybek
English review Renée Gadsden
Graphic art and design Milojka Žalik Huzjan
Cover image Jurij Fikfak

Published by Institute of Slovenian Ethnology at ZRC SAZU
ZRC Publishing, Ljubljana
Österreichische Gesellschaft für Semiotik – Institut für Sozio-Semiotische Studien, Vienna
Represented by Monika Kropej, Oto Luthar, Jeff Bernard
Co-publishers Department of Slavic Studies, Graz University
Institutum Studiorum Humanitatis, Ljubljana
Editor-in-Chief Vojislav Likar
Printed by Collegium graphicum, d. o. o., Ljubljana

The publication and symposium were directly or indirectly supported by
Slovenian Research Agency, Ljubljana; Graz University (Vice Rector and Office for International Relations, Faculty for Cultural Studies, Department of Slavic Studies); Office of the Government of the Province of Styria (Department for Science); Office of the Mayor of the City of Graz; Austrian Science and Research Liaison Office Ljubljana (ASO); Austrian Federal Ministry of Education, Science and Culture, and the City of Vienna.
The organizers and co-editors express their warm thanks for these valuable efforts.

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

81'27(082)

TEXT & reality = Text & Wirklichkeit / edited by Jeff Bernard, Jurij Fikfak, Peter Grzybek. - Ljubljana : Institute of Slovenian Ethnology at ZRC SAZU, ZRC Publishing : Institutum Studiorum Humanitatis ; Vienna : Österreichische Gesellschaft für Semiotik, Institut für Sozio-Semiotische Studien ; Graz : Department of Slavic Studies, University, 2005

ISBN 961-6500-86-4 (ZRC SAZU)
ISBN 3-900494-46-0 (Österreichische Gesellschaft für Semiotik)
1. Vzp. stv. nasl. 2. Bernard, Jeff
220472576

TEXT & REALITY
TEXT & WIRKLICHKEIT

EDITED BY: JEFF BERNARD
JURIJ FIKFAK
PETER GRZYBEK

LJUBLJANA, WIEN, GRAZ 2005

CONTENTS