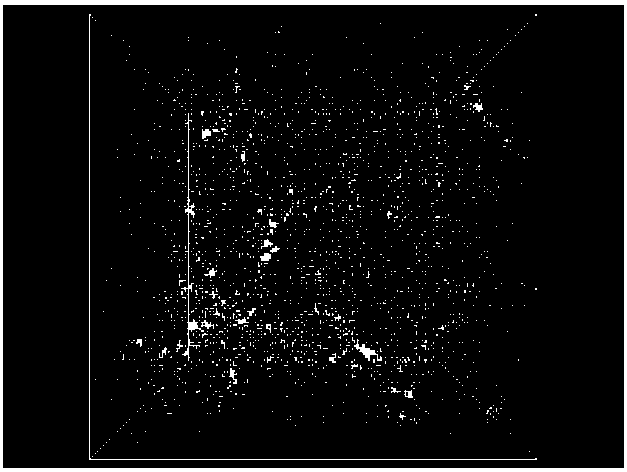
Peter Grzybek/Emmerich Kelih:

Textforschung: Empirisch !



0. Texte als Objekte der Textwissenschaft

Wenn wir davon ausgehen, dass Wissenschaft sich nicht (nur) mit der Beschreibung einzelner Objekte begnügen darf, sondern nach Regularitäten und Gesetzmäßigkeiten suchen muss sowie nach theoretischen Zusammenhängen, was in letzter Konsequenz auf eine Erklärung der Phänomene hinausläuft, dann darf sich eine Wissenschaft der Texte nicht mit der Beschreibung einzelner Texte zufrieden geben. Vielmehr muss es in einem ersten Schritt darum gehen, die bestimmten Texten gemeinsamen Eigenschaften aufzufinden und beschreibbar zu machen, d. h. eine bestimmte Art der Kategorisierung zu erreichen. In einem zweiten Schritt müssen dann die Wechselbeziehungen zwischen den verschiedenen Eigenschaften der einzelnen Texte und der Textkategorien so erfasst werden, dass diese Wechselbeziehungen gesetzmäßig formuliert werden können, so dass wir letztendlich zu einer Theorie der Texte und Textkategorien sowie ihres Funktionierens in der Kommunikation gelangen.

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 Textsortenforschung. Während es dabei in den traditionellsten Formen vor allem der Literaturwissenschaft lange um die Gegenüberstellung von formalen und inhaltlichen Kriterien ging, hat sich nicht zuletzt unter dem Einfluss der Sprachwissenschaft eine funktionale Sichtweise dieser Elemente durchgesetzt, welche die textinternen funktionalen Aspekte in den Vordergrund stellte. Ebenfalls unter dem Einfluss kommunikationstheoretischer und pragmatischer Fragestellungen hat sich dann die Funktionalstilistik herausgebildet, der es im Prinzip darum ging, stilistische Merkmale auf der textinternen Ebene mit pragmatischen Funktionen auf der

textexternen Ebene zu verbinden. In dieser Tradition steht in letzter Konsequenz – bei allen Unterschieden – auch die sehr viel aktuellere Textsortenforschung, die Textsorten als Klassen von Texten versteht, die von gemeinsamen inhaltlichen („thematisch-propositionalen"), stilistischen und handlungstypisch-illokutiven Grundelementen bestimmt sind. Beide Herangehensweisen stehen in gewisser Weise an zwei unterschiedlichen Enden, insofern es der als top-down-orientierten Funktionalstilistik um eine maximale Reduktion (d.h. um eine minimale Anzahl) von Textklassen ging, während es der sich selbst als empirisch bezeichnenden Textsortenforschung um eine maximale Ausdifferenzierung des textkommunikativen Geschehens geht. Dennoch haben beide Vorgangsweisen als entscheidendes Merkmal gemeinsam, dass sie **ausschließlich qualitativ** ausgerichtet sind. In beiden Fällen wird sozusagen die „Welt der Texte" in dem Sinne mit Bezug zur Welt strukturiert, als sie unter Heranziehung von textexternen (pragmatischen) Faktoren strukturiert werden soll: Während es bei den Funktionalstilen um allgemeine kommunikative (gesellschaftlich definierte) Sprach- und/oder Textfunktionen geht, handelt es sich bei der Textsortenforschung um spezifische kommunikativ-situative Funktionen. Damit verbunden ist natürlich eine extreme Unterschiedlichkeit in der Anzahl der resultierenden Kategorien: Während die Funktionalstilistik sich – je nach konkreter Schule – in der Regel mit der Unterscheidung von ca. fünf bis acht verschiedenen Funktionalstilen begnügte, hat die Textsortenforschung es auf ein Inventar von nicht weniger als ca. 4000 verschiedenen Textsorten gebracht.
Es stellt sich in diesem Zusammenhang natürlich die Frage, inwiefern beide Ansätze miteinander kompatibel sind. Mit anderen Worten: Ist es möglich, Textsorten mit einem hohen Grad an intersubjektiver Übereinstimmung auf einer höheren Ebene zu Gruppen, zu bestimmten Diskurstypen oder Funktionalstilen zuzuordnen? Dieser Frage soll im vorliegenden Beitrag nachgegangen werden, und zwar auf zweierlei Art und Weise – in beiden Fällen empirisch. Wir verstehen unseren Ansatz somit als einen Beitrag zur empirischen Textforschung.

1. Textforschung empirisch (I): subjektbezogen

In einer Untersuchung mit Leipziger Studierenden der Germanistik, die sowohl mit den Grundlagen der Funktionalstilistik als auch der Textsortenforschung vertraut waren, sollte herausgefunden werden, ob und in welchem Maße intersubjektive Übereinstimmung bei der texttheoretischen Zuordnung von Textsorten zu Funktionalstilen erreicht werden kann. Zugrunde gelegt wurde dabei zwei Schemata:

(a) ein Schema der Funktionalstile, wie es in Abb. 1 dargestellt ist; im Bewusstsein, dass es in der Geschichte der Funktionalstilforschung durchaus verschiedene Einteilungen gegeben hat, orientiert sich das in Abb. 1 dargestellte Schema im wesentlichen an dem in der tschechoslowakischen Tradition stehende von Mistrík (1973)

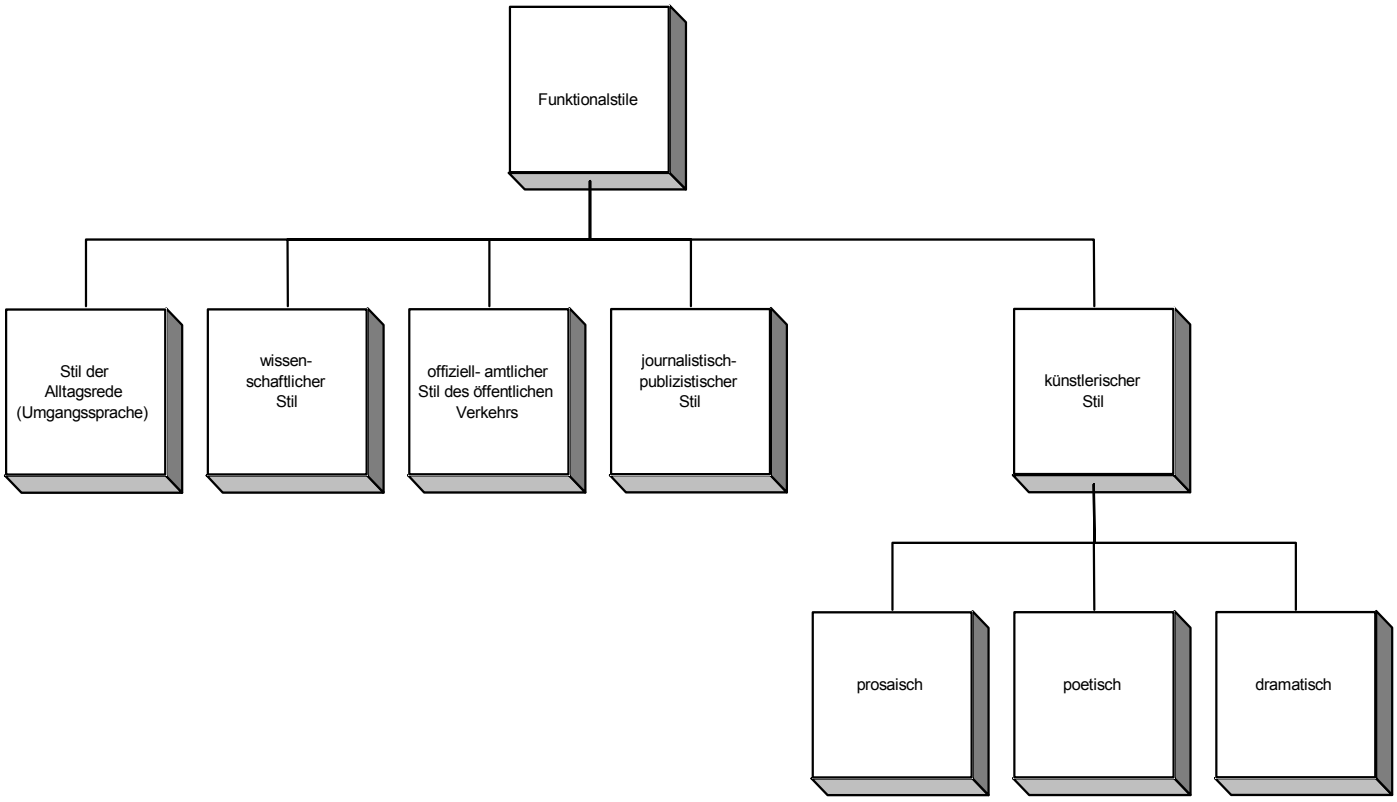


Abb. 1: Funktionalstile nach Mistrík (1973)

(b) ein Schema ausgewählter Textsorten, wie es in Tab. 1 zu sehen ist; die Auswahl wurde dabei so motiviert, dass in einer ersten (subjektiven und tentativen) Zuordnung möglichst das gesamte Spektrum der Funktionalstile mit verschiedenen Textsorten repräsentiert ist.

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

Tab.1: (Subjektive und tentative) Zuordnung von Textsorten zu Funktionalstilen

Den befragten Personen wurde allerdings keines der beiden Schemata präsentiert, vielmehr nur eine alphabetisch sortierte Liste mit der Bezeichnung der 67 Textsorten aus Tab. 1 sowie der folgenden Zahlenkodierung:

Funktionalstile

1 Alltag / privat

2 Wissenschaft

3 Administration / öffentlicher Verkehr

4 Journalistik

5 Prosa

6 Poesie } künstlerischer Stil

7 Dramatik

Dabei sollten die Befragten jeder Textsorte eine dieser Zahlen (1 bis 7) zuordnen; für den Fall, dass eine Person eine Textsorte mehr als einem Funktionalstil (aber nicht mehr als zwei verschiedenen) zuordnen wollte, wurde eine zweistellige Codierung vorgesehen (also z. B. „14“ für eine Kombination von Alltag/privat und Journalistik).
Im Ergebnis stellte sich heraus, dass von den 67 Textsorten nur vier (ca. 6%) von den Befragten mit 100%iger Übereinstimmung ein und demselben Funktionalstil zugeordnet wurden. Hierbei handelt es sich um die Textsorten „Diplomarbeit“ und „Dissertation“ (2 = Wissenschaft) einerseits, um die Textsorten „Gedicht“ und „Sonett“ (6 = künstlerischer Stil - Poesie) andererseits. Die Anzahl vollständiger Übereinstimmung erhöht sich auf 13 (19,40%), wenn man auch die Doppelzuordnungen berücksichtigt und zur Zuordnung zu einer Textsorte hinzurechnet. Das heißt aber, dass es bei 80% der Textsorten keine vollständige Übereinstimmung in der Zuordnung zu Funktionalstilen gibt.
Diese Einschätzung ist natürlich insofern mit Vorsicht zu genießen, als es hier nur um 100% Übereinstimmung geht. Unter Einschluss der Doppelzuordnungen lässt sich immerhin bei 39 der 67 Textsorten (also 58,21%) eine mehr als 90%-ige Übereinstimmung, bei 59 der 67 Textsorten (88,06%) eine Übereinstimmung bei mehr als zwei Drittel der Befragten erreichen.
Angemessener scheint es jedoch auszurechnen, wie hoch der Grad der Übereinstimmung insgesamt im Durchschnitt ist. Notwendig dazu ist die Berechnung für jede einzelne der 67 Textsorten, welcher Funktionalstil jeweils am meisten Zuordnungen auf sich vereinigen konnte, wie hoch dieser Prozentsatz ist und wie hoch dann der Grad der Übereinstimmung

im Durchschnitt ist. Die einzelnen Ergebnisse sind in der Tab. 3 (s. u.) detailliert dargestellt – im Durchschnitt kommt eine Textsorte auf eine Übereinstimmung von 67,12%. Dies gilt, wenn man nur die Einfachzuordnungen berücksichtigt; addiert man hier auch die entsprechenden Sekundärzuordnungen hinzu, ist der Grad an Übereinstimmung mit 86,35% insgesamt relativ hoch. Interessant ist auch der Vergleich mit der angenommenen Voraus-Klassifikation: Denn in der Tat führen die Ergebnisse dazu, dass in 90% der Fälle die Textsorten mehrheitlich zu genau den Funktionalstilen zugeordnet werden, wie dies auch in dem in Tab. 1 angegebenen Schema vorgeschlagen wurde.
Damit lässt sich die Schlussfolgerung ziehen, dass insgesamt ein nicht geringer Grad an intersubjektiver Übereinstimmung erzielt werden kann, wenn es darum geht, einzelne Textsorten bestimmten Funktionalstilen zuzuordnen. Wenn man nur eindeutige Zuordnungen zulässt, kommt man dabei zwar nur auf einen Übereinstimmungsgrad von ca. 67%; wenn man allerdings auch Doppelklassifikationen mit in die Auswertung einbezieht – was bei einer Anzahl von sieben Funktionalstilen natürlich den Wert der Aussage stark mindert – erhöht sich der Grad auf ca. 85%.
In der subjekt-bezogenen Text-Empirie würde man es mit diesen Ergebnissen auf sich beruhen lassen und könnte mit ihnen – mehr oder weniger (?) – zufrieden sein. Allerdings stellt sich das Ergebnis aus der Perspektive objekt-bezogener Empirie vollkommen anders dar. Aus dieser Sichtweise nämlich handelt es sich bei diesen Ergebnissen der Zuordnung lediglich um qualitative Apriori-Klassifikationen, die im Weiteren einer empirischen Überprüfung am konkreten Text-Material zu unterziehen sind. Ausgangspunkt dieser Sichtweise ist die Annahme, dass sich die Texte selbst durch spezifische sprachlich-textuelle Merkmale auszeichnen, die ihrerseits eine Klassifikation und Typologie erlauben. Um die Reichweite dieser Annahme zu skizzieren, ist es notwendig, einleitend einen etwas weiteren theoretischen Bogen zu schlagen, der den Horizont dieser im Bereich der Textwissenschaft eher unüblichen Herangehensweise skizziert.

2. (Text-)Theoretische Grundlagen

Wenn wir einmal annehmen – und zwar in Form einer wissenschaftlichen Analogie, nicht etwa nur eines metaphorischen Vergleichs –, dass wir es mit einem *Universum von Texten* zu tun haben, d. h. mit einer endlichen oder aber unendlichen Anzahl textueller Objekte, die ein offenes oder geschlossenes System verkörpern, dann stellt sich als erstes die Frage, ob dieses Universum strukturiert ist bzw. strukturiert werden kann – und wenn ja: wie. Unter der Annahme der Existenz einer solchen Struktur – zumindest als anzustrebendes Ergebnis der wissenschaftlichen Erkennt-

nis – muss eine Beschreibung des Universums zwei verschiedene Prozesse beinhalten, die notwendigerweise ineinander greifen müssen:

(a) die Identifikation der Objekte des Systems (die im gegebenen Fall auf einer Definition von ‚Text' beruht),

(b) die Klassifikation dieser Objekte, die in der Identifikation und Beschreibung hierarchisch geordneter (im gegebenen Fall textueller) Subsysteme resultiert.

Unter Beibehaltung der astronomischen Analogie – die wir, wie gesagt, keineswegs nur als Bild oder Metapher, sondern als wissenschaftstheoretisches Postulat verstehen – 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 Astral- oder Sonnensystemen zu sehen sind.
Es ist selbstverständlich, dass die beiden Prozesse der Identifikation und Klassifikation nicht ohne Rückgriff auf bestimmte 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 letztendlich jede Art von Klassifikation auf die eine oder andere Art und Weise und zu unterschiedlichem Maße quantifizierende Vorgangsweisen. Auch scheinbar rein qualitative Herangehensweisen (die in den sich mit Texten beschäftigenden Disziplinen fraglos überwiegen) kommen nicht ohne quantitative Argumente aus – und sei es nur in Form eines implizit oder explizit enthaltenen Postulats, dass die einen Objekte einander „mehr oder weniger" ähnlich seien, weil sie „mehr oder weniger" Eigenschaften gemein haben, weil sie in „größerer oder kleinerer" Nähe oder Entfernung zueinander stehen, weil sie in „höherem oder geringerem" Maße einer vermeintlichen Norm oder einem Prototyp entsprechen usw. Der Grad der Quantifizierung wird dabei von den in der jeweiligen Meta-Sprache enthaltenen Eigenschaften bestimmt. Deswegen ist es von besonderer Bedeutung, von welcher analytischen Ebene der Klassifikationsprozess ausgeht, wobei freilich jede eigene Ebene mit jeweils spezifischen Problemen behaftet ist, was die Definition der Sub-Systeme und deren Grenzen betrifft.
Ungeachtet dieser Probleme lässt sich die beschriebene Klassifikation des (Text)-Universums nicht ohne empirische Methoden erreichen. Im Zusammenhang mit dieser pauschalen Aussage darf man freilich nicht übersehen, dass Auffassung und Verständnis von empirischer Arbeit in verschiedenen Wissenschaftsdisziplinen sehr unterschiedlich aussehen,

seien diese nun mit sprachlichen Objekten befasst oder nicht. Auch differiert der Anteil von Theorie und Praxis, die Gewichtung quantitativer und qualitativer Argumente, von Disziplin zu Disziplin beträchtlich, wobei Disziplinen, die sich auf die eine oder andere Art und Weise mit sprachlichen Objekten beschäftigen, in der Regel eher dazu neigen, theoretische und qualitative Zugangsweisen zu favorisieren. Als stark empirisch ausgerichtete Disziplin hat sich neben diesen „traditionellen" Methoden in den vergangenen Jahren insbesondere die sog. Korpus-Linguistik als eine spezifische Disziplin bzw. Sub-Disziplin etabliert, die mit Sprache(n) und Text(en) operiert. Sich selbst als eine „daten-orientierte" Disziplin verstehend, lautet die Grundannahme der Korpuslinguistik, dass eine Maximierung der Datenbasis zu einer zunehmend „repräsentativen" Beschreibung von Sprache führt.
Keine dieser Disziplinen – seien sie nun überwiegend theoretisch oder empirisch ausgerichtet – kommt jedoch letztendlich ganz ohne quantitative Aussagen oder Methoden aus. Deshalb kommt an dieser Stelle die quantitative Linguistik als eine wichtige sprachwissenschaftliche Disziplin ins Spiel: Im Gegensatz zu den oben beschriebenen linguistischen 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 letztendliche Interpretation der Ergebnisse in einen theoretischen Rahmen eingebettet.
In seinem Bemühen um eine quantitativ fundierte Texttypologie folgt der hier von uns vorgeschlagene Ansatz diesen allgemeinen Grundüberlegungen. Im Gegensatz zu der erwähnten Teil-Disziplin der Korpuslinguistik ist unser Ansatz – der sich am ehesten als **quantitative Textanalyse** bezeichnen lässt – von zwei Grundgedanken gekennzeichnet: Abgesehen von der überwiegend theoretischen Ausrichtung ist dies die Annahme, dass ein ‚Text' die relevante analytische Einheit ist, die einer Untersuchung zugrunde liegt. Insofern es der Korpuslinguistik um die Konstruktion bzw. Re-Konstruktion bestimmter Normen, „repräsentativer" Standards o. Ä. der (oder einer) Sprache geht, arbeiten korpus-basierte Analysen in der Regel mit Mischungen heterogener Texte, die in einem sich aus dieser Sicht als „Quasi-Text" darstellenden Korpus zusammengeführt werden. Im Gegensatz dazu richtet sich die quantitative Textanalyse auf individuelle **Texte als homogene Einheiten**. Die Grundannahme lautet, dass ein (vollständiger) Text ein selbst-regulierendes System ist, das von spezifischen Regularitäten organisiert wird. Diese Regularitäten müssen nicht zwangsläufig auch in Textsegmenten vorliegen, und es ist wahrscheinlich, dass sie sich in jeglicher Art von Textkombination überschneiden (und damit gegenseitig überlagern oder neutralisieren).

Mit der Annahme, dass Gegenstand der Analyse homogene Texte sein sollten, ist natürlich noch keine Definition von ‚Text' geliefert – insofern 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 allgemein gültige Antwort auf diese Frage in den verschiedenen Textwissenschaften, und so taucht die Frage, ob es eines „neuen" Textbegriffs bedarf, mit schöner Regelmäßigkeit immer wieder in den textwissenschaftlichen Debatten auf.
In der vorliegenden Darstellung kann es nicht um eine theoretische Lösung dieser Frage gehen. Aus unserer Sicht, die einen spezifischen (nicht zuletzt auch statistisch geprägten) Blick auf das Problem wirft, stellt sich die Problematik des Sachverhalts in Form von zweierlei Teilproblemen dar:

(i) das Problem der Datenhomogenität,

(ii) das Problem der zugrunde gelegten 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 auf einzelne 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 (vgl. dazu Altmann 1992 und Orlov 1972). Die für die einzelnen Texte erhaltenen Beschreibungsmodelle lassen sich in weiterer Folge miteinander vergleichen, was möglicherweise (erstrebenswerterweise) in einem oder mehreren allgemeinen Modellen resultiert; auf diese Weise lässt sich eine quantitative Texttypologie erreichen.
2. Auch mit einer bestimmten Definition von ‚Text' bleibt zu entscheiden, welche Text-Eigenschaften einer quantitativen Analyse unterzogen werden sollen; in der vorliegende Studien konzentrieren wir uns auf die Wortlänge als einer spezifischen Texteigenschaft, in vollem Bewusstein, dass Wortlänge nur eine von mehreren, zudem mit anderen stark interagierende Text-Eigenschaft darstellt.

3. Wortlänge und quantitativ-qualitative Methoden der Textforschung

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. Wortlänge ist dabei zu verstehen als eine Möglichkeit, einen wesentlichen Teilaspekt der Struktur von Texten quantifizierend zu beschreiben und für weiterführende Fragestellungen, wie etwa die Typologisierung von Texten, heranzuziehen. Wortlänge ist dabei nicht als das einzige Maß zu verstehen, welches für derartige Fragestellungen heranzuziehen ist. Vorstellbar und anwendbar ist ein ganzes Set von quantitativen Eigenschaften eines Textes (Satzlänge, Phrasenlänge, Häufigkeiten verschiedener Texteinheiten wie lexikalischer Reichtum, Textdeckung, das lexikalisches Type-Token-Verhältnis u. a. m.), wobei jede Texteigenschaft allein oder aber in (synergetischer) Beziehung zu einer oder mehreren anderen Eigenschaften betrachtet werden kann.
In jedem Fall ist für diese Beschreibungsebene die Anwendung von a priori festzulegenden Definitionen der Texteinheiten notwendig, die einer Quantifizierung unterzogen werden. Im vorliegenden Fall wird für die Zwecke einer automatischen Textanalyse das Wort als eine graphematische Einheit definiert – zur Anwendung von anderen Definitionen vgl. Antić/Kelih/Grzybek (2004) –, die sich in elektronisch vorliegenden Texten als eine durch Leerstellen abgrenzte Einheit definieren lässt. Anzumerken ist, dass die Texte vor der automatisierten Bestimmung der Wortlänge einheitlich bearbeitet und vor der automatischen Analyse einer spezifischen Tagging-Prozedur unterzogen werden: So werden z. B. vorkommende Abkürzungen, Zahlen, Eigennamen u. Ä. in der entsprechenden grammatikalisch-morphologischen Form aufgelöst, in Privatbriefen werden Datum und Adresse u. Ä. ausgeklammert, in den Dramen werden Regieanweisungen u. Ä. nicht in die Untersuchung einbezogen, usw.). In den Texten werden alle vorkommenden Wörter bzw. Wortformen (d. h. Tokens) gezählt.
Die Wortlänge lässt sich im Prinzip in verschiedenen Maßeinheiten berechnen, sei es, dass sie in der Anzahl der Buchstaben (Grapheme), der Anzahl der Phoneme, der Morpheme oder der Anzahl der Silben pro Wort bestimmt werden. Es liegt auf der Hand, dass die Bestimmung der Maßeinheit (bzw. die Entscheidung für eine bestimmte Maßeinheit) nicht ohne Auswirkung bleibt, wobei natürlich insbesondere systematische Verschiebungen, die durch die Wahl unterschiedlicher Maßeinheiten bedingt werden, von besonderem Interesse sind (vgl. Grzybek/Kelih 2004b). Die im Ergebnis erhaltenen Rohdaten (d. h. die Anzahl *x*-silbiger

Wörter in einem Text) werden in weiterer Folge in eine sog. Wortlängenhäufigkeitsverteilung transformiert. Abb. 2 stellt graphisch ein Beispiel einer solchen Verteilung dar, nämlich die Anzahl der *x*-silbigen Wörtern in einem slowenischenText; dabei handelt es sich um einen Zeitungskommentar aus dem unten noch näher zu beschreibenden Text-Korpus.

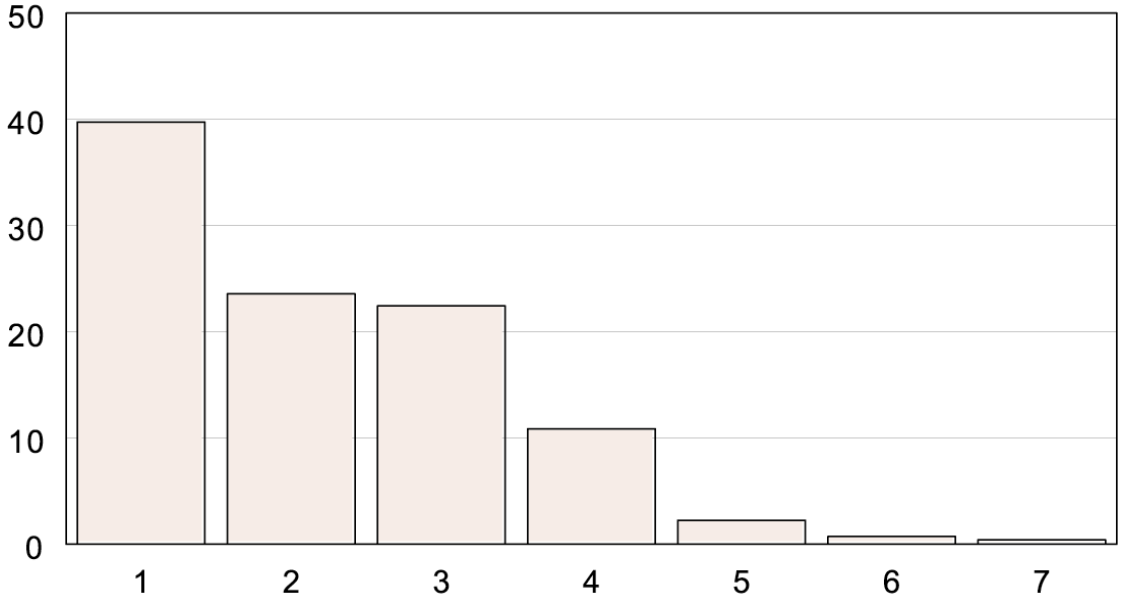


Abb. 2: Häufigkeitsverteilung der x-silbigen Wörter in einem Zeitungskommentar

Die Häufigkeitsverteilung der *x*-silbigen Wörtern ist sodann der Ausgangspunkt für die weitere Berechnung von statistischen Kenngrößen. Die allgemein vermutliche 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, Kurtosis, Entropie, Wiederholungsrate und andere Kenngrößen sind in der Sprach- und Textwissenschaft wiederholt zur Anwendung gekommen. Diese und weitere Kenngrößen – im Grazer Wortlängen-Projekt wurde ein Set von nicht weniger als 30 solcher Kenngrößen erarbeitet – lassen sich sodann als Variablen in den Analysen verwenden. Nicht immer und nicht für alle Fragestellungen ist es notwendig, die Gesamtheit der theoretisch zur Verfügung stehenden Kenngrößen zum Einsatz zu bringen; oft reichen zwei oder drei von ihnen aus, um eine gegebene Fragestellung in befriedigendem Maße zu lösen. Bei der Analyse bemüht man sich natürlich, die Anzahl der verwendeten Kenngrößen möglichst gering zu halten, damit die getroffene Aussage um so leichter interpretierbar wird (womit die quantitative Herangehensweise in eine als komplementär zu verstehende qualitative überführt wird). Für die vorliegende Fragestel-

lung einer quantitativen Texttypologisierung etwa wird sich ein Set von maximal vier Variablen als ausreichend erweisen; in der Regel reichen jedoch zwei oder drei Variablen aus, um eine optimale Klassifizierung zu gewährleisten – doch dazu unten mehr.
Hat man nun sowohl die Daten der Häufigkeitsverteilungen und die entsprechenden Kenngrößen zu jedem einzelnen Text, dann lassen sich die Werte in entsprechende Spezialprogramme eingeben und statistisch bearbeiten. Die zur Verfügung stehenden Möglichkeiten und Methoden können hier nicht im einzelnen erörtert werden, dennoch sollte das Prinzip der Optionen deutlich gemacht werden – wobei die traditionellen qualitativen Methoden hier nicht mehr zur Sprache kommen müssen, und wir uns auf die quantitativ-qualitativen konzentrieren können:

1. Anwendbar ist prinzipiell eine solche quantitative Klassifizierung von Texten, die zunächst (!) keine qualitativen (den Texten zugeschriebenen) Merkmale in die Analyse einführt, sondern ausschließlich quantitative Informationen, die im gegebenen Fall in Form der Wortlänge und daraus abgeleiteter Kenngrößen vorliegen. Derartige Verfahren – die sich bedingt als auf einem **„Tabula-Rasa-Prinzip“** basierende Methoden verstehen lassen – beinhalten z. B. die Anwendung von so genannten Cluster-Analysen und post-hoc-Analysen (s. u.).
2. Ein zweite Herangehensweise lässt sich bedingt als „**A-priori / A-posteriori-Prinzip“** bezeichnen, insofern diese eine Synthese der Textklassifizierung auf der Grundlage von qualitativen und quantitativen Merkmalen darstellt. Dabei werden im gegebenen Fall die Textsorten zunächst (a priori) tentativ bestimmten Funktionalstilen zugeordnet (vgl. dazu die Tab. 1); a posteriori werden die Texte incl. dieser Zuordnungen aufgrund der aus der Wortlänge erhaltenen statischen Kenngrößen multivariaten Diskriminanzanalysen unterzogen. Damit kann empirisch getestet werden, inwiefern (d. h. in welchem Maße) die qualitative Zuordnung dem quantitativen Informationsbestand der untersuchten Texte entspricht.

4. Die Texte der vorliegenden Untersuchung

Wie einleitend erwähnt, geht es in der hier vorgestellten Studie exemplarisch um die Analyse von 398 slowenischen Einzeltexten; Detailangaben zur Struktur dieses Textkorpus sind in Tab. 2 aufgeschlüsselt:

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. 2. Textbasis: 398 slowenische Texte

Die Texte haben den Status jeweils in sich abgeschlossener Einzeltexten, wobei die einzelnen Kapitel eines Kurzromans jeweils als eigene ‚Texte' definiert werden – Grund für diese Art der Textdefinition ist, wie oben bereits erwähnt wurde,

(a) die ausgängliche Annahme, dass es sich hierbei um das Ergebnis jeweils relativ homogener Prozesse der Textgenierung handelt, und

(b) die weiterführende Annahme, dass (nur) innerhalb der so definierten Texte bestimmte Prozesse sprachlich-textueller Selbstregulation wirksam sind.

Deutlich zu sehen ist, dass dabei ein Fokus auf verschiedenen Briefsorten unterschiedlicher Funktionalstile liegt (Privatbriefe, Offene Briefe, Leserbriefe, Briefe aus einem Briefroman). Diese spezifische Selektion ist nicht zufällig motiviert: Abgesehen davon, dass Briefe wohl am ehesten als Ergebnis eines homogenen Produktionsprozesses sind, geht man bei ihnen davon aus, dass sie insofern eine für Sprach- und Textuntersuchungen prototypische Textsorte darstellen (vgl. Köhler 1998, ii), als sie an der Schnittstelle zwischen Schriftlichkeit und Mündlichkeit liegen (bzw. Elemente von beidem in sich vereinigen).
Insgesamt weist das Text-Korpus somit nicht weniger als 398 in die Analysen eingehende Texte auf, die acht verschiedenen Textsorten entsprechen. Um einen Vergleich mit den oben dargestellten subjekt-bezogenen Zuordnungen zu ermöglichen, führt Tab. 3 die Ergebnisse der Befragung für diese acht Textsorten an.

		n	%
Privatbrief	1_Alltag	21	88
	13	1	4,2
	15	1	4,2
	16	1	4,2
	Gesamt	24	100

		n	%
Geschäftsbrief	3_Öffentlich	23	95,8
	31	1	4,17
	Gesamt	24	100

		n	%
Leserbrief	1_Alltag	8	33
	4_Journalistik	5	21
	13	1	4,2
	14	8	33
	15	1	4,2
	34	1	4,2
	Gesamt	24	100

		n	%
Briefroman	1_Alltag	1	4,17
	5_Kunst-Prosa	16	66,7
	6_Kunst-Poesie	1	4,17
	15	5	20,8
	56	1	4,17
	Gesamt	24	100

		n	%
Kommentar	3_Öffentlich	2	8,3
	4_Journalistik	15	63
	14	2	8,3
	24	1	4,2
	41	2	8,3
	42	1	4,2
	45	1	4,2
	Gesamt	24	100

		n	%
Kurzroman	5_Kunst-Prosa	21	87,5
	15	1	4,2
	45	1	4,2
	56	1	4,2
	Gesamt	24	100

		n	%
Drama	7_Kunst-Drama	21	88
	17	2	8,3
	76	1	4,2
	Gesamt	24	100

		n	%
Gedicht	6_Kunst-Poesie		
		24	100

Tab. 3: Ergebnisse der empirischen Umfrage zur Zuordnung von Textsorten zu Funktionalstilen

5. Quantitativ-Qualitative Analysen

Im Sinne des oben beschriebenen Unterschieds zwischen „rein quantitativen“ (d. h. ohne qualitative Apriori-Annahmen funktionierenden) und „quantitativ-qualitativen“ Methoden lässt sich sagen, dass mit Hilfe von **Cluster-Analyse** versucht wird, gegebene Fälle (hier: Texte) anhand von gegebenen Variablen (hier: Wortlänge und weitere daraus abgeleitete Kenngrößen) so in Gruppen (d.h. Cluster) einzuteilen, dass die Fälle eines Clusters bezüglich dieser Variablen ähnliche Werte und die Fälle aus verschiedenen Clustern möglichst unähnliche Vari-ablenwerte aufweisen. Wie die einem Cluster zugeordneten Fälle sich qualitativ darstellen, ist dabei zunächst einmal nicht von Interesse. Es würde über den Rahmen des hier Möglichen hinausgehen, weitere Details zu erörtern (vgl. Grzybek et al. 2004); deshalb sei nur resümierend erwähnt, dass die 398 Texte in einem ersten Schritt einer entsprechenden Cluster-Analyse unterzogen wurden. Die entsprechenden angewandten Verfahren (Ellbow-Technik und Two-Step-Analysen mit dem Log-Likelihood-Distanzmaß) zeigen, dass es sich im Hinblick auf eine optimale Anzahl von Clustern anbietet, das Textmaterial aufgrund der quantitativen Information zur Wortlänge in **drei Cluster** zu untergliedern. Das würde bedeuten, dass eine Zusammenfassung der acht Textsorten in drei Obergruppen (ob dies nun Funktionalstile sind oder nicht) einer optimalen Gruppenzahl von drei Gruppen gleichkäme.
Im Gegensatz zu den oben angesprochenen Cluster-Analysen führen sog. **Post-hoc-Analysen** durchaus qualitative Merkmale mit in die Analyse ein. Meistens werden sie in Form von Post-hoc-Mittelwertvergleichen durchgeführt, sie lassen sich jedoch im Prinzip auf alle anderen Variablen anwenden. Post-hoc-Analysen zielen (im gegebenen Fall) auf eine Beantwortung der Frage, welche Textsorten sich so zusammenfassen lassen, dass „homogene Untergruppen“ gebildet werden. Entsprechenden Analysen der mittleren Wortlänge in unseren acht Textsorten führen zur Unterscheidung von fünf homogenen Untergruppen (vgl. Tab. 4).

Funktionalstil	Untergruppen für α = 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				2,883	
Offene Briefe					2,4268

Tab. 4: Resultate der post-hoc-Analysen

Allgemein widerspricht diese Beobachtung, dass fünf Untergruppen unterschieden werden, dem Ergebnis der Clusteranalyse, nach der die Zusammenfassung aller Texte in insgesamt drei Gruppen eine optimale Kategorisierung wäre. Das kann nur bedeuten, dass sich einige der Textsorten nicht (bzw. nicht allein) aufgrund der Wortlänge voneinander differenzieren lassen, sondern in eine gemeinsame übergeordnete Gruppe eingehen. Theoretisch wäre es möglich, dass sich hier eine Zuordnung der Textsorten zu den Funktionalstilen ergibt – möglich wäre aber auch, dass die Obergruppen von ganz anderer Qualität sind und in einer neuen Art von Typologie resultieren.
Als interessante Beobachtung ist in jedem Fall zu bemerken, dass als Ergebnis der Post-hoc-Analyse die vier verschiedenen Brieftypen in vier unterschiedliche Untergruppen fallen. Insofern ist davon auszugehen, dass der Typus Brief in sich nicht homogen ist, sondern ein relativ breites Spektrum abdeckt.
Aufschluss darüber geben können hier nur **multivariate Diskrimanzanalysen:** 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 Prädikatorvariablen (also unseren Kenngrößen) bestimmten Gruppen zuzuordnen; dabei werden die Variablen so transformiert, dass im Ergebnis eine optimale Diskrimination der einzelnen Fälle herauskommt: Es lässt sich dann also sagen, wie viele der Fälle (bzw. wie viel Prozent der Fälle) aufgrund der verwendeten Kenngrößen „richtig“ den a priori zugeordneten Kategorien zugeordnet werden.
In einem ersten Schritt wird das gesamte ’Korpus’ der slowenischen Texte – jeweils tentativ einer der acht Textsorten zugeordnet – einer derartigen Diskriminanzanalyse unterzogen. Auf der Basis von zwei Prädikatorenvariablen – dem arithmetischen Mittel und dem relativen Anteil von

einsilbigen Wörtern – ergibt sich eine korrekte Zuordnung von nicht mehr als 56,30% der Texte. Dieser Sachverhalt ist in Abb. 3 anschaulich dargestellt.

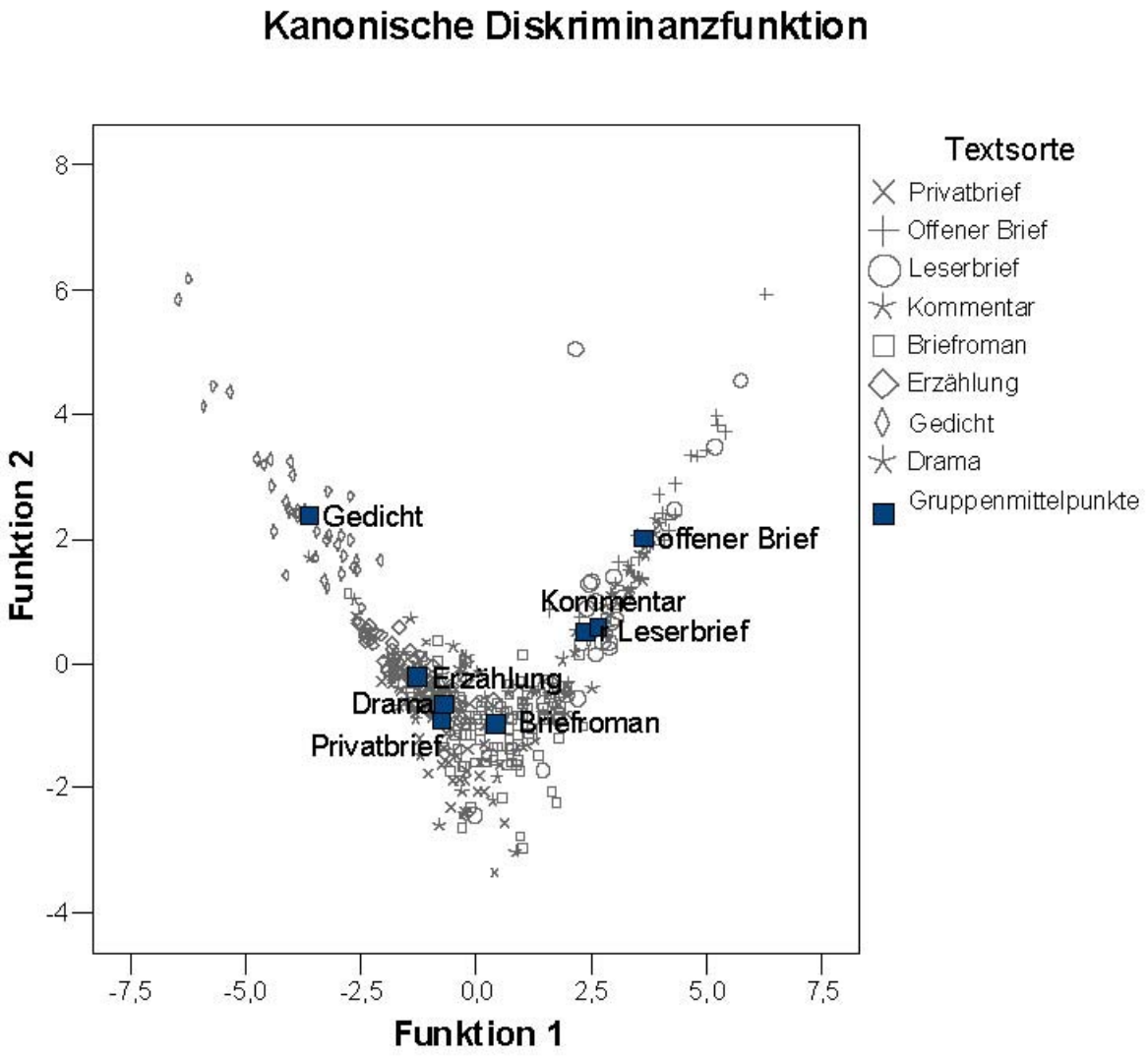


Abb. 3: Diskriminanz-Analyse (8 Textsorten)

Folgende Tendenzen sind klar zu erkennen: Alle Textsorten nehmen jeweils ein mehr oder weniger klar abgegrenztes Feld ein. Zu beobachten ist jedoch auch eine ganze Reihe von Überlappungen der einzelnen Textsorten, wobei vor allem die versgebundenen Gedichte einen klar abgegrenzten Bereich einnehmen; ähnliches gilt auch für die Leserbriefe, die offenen Briefe und die Kommentare. Die dramatischen Texte, die Erzählungen, die Privat-Briefe und die Briefe aus dem Briefroman nehmen im Gegensatz dazu ein etwas weiteres, nicht exakt abgrenzbares Feld ein.

Dieses erste Ergebnis führt zu der Überlegung, in einem ersten Schritt die gesamte Textbasis vorübergehend durch die Elimination einiger Textsorten zu reduzieren, damit bestimmte Grundmuster des Textuniversums gegebenenfalls transparenter erscheinen. In einem zweiten Schritt wären dann die einzelnen Textsorten sukzessiv wieder einzuführen; dabei wären jeweils nach Wiedereinführung einer Textsorte Diskriminanzanalysen durchzuführen, welche durch den Prozentsatz der korrekt diskriminieren Texte direkt Auskunft über die Effizienz der Zuordnung zu einer der bestehenden Gruppen gibt. Letztendlich sollte so eine optimierte Zuordnung aller Texte bzw. Textsorten zu übergeordneten Gruppen erreicht werden.
Im Hinblick auf den ersten Schritt bietet es sich an, das gesamte Textmaterial zunächst auf die vier Brieftypen (Leserbriefe, offene Briefe, Briefroman, Privatbriefe) zu reduzieren, die ja bei den oben erwähnten post-hoc-Analysen vier unterschiedlichen homogenen Untergruppen zugeordnet wurden. Belässt man es bei diesen vier Textsorten, die durch immerhin 213 Texte repräsentiert sind, resultiert dies in einem relativ niedrigen Prozentsatz von 70,4% korrekt diskriminierter Texte. Dabei zeigt sich allerdings, dass insbesondere Privatbriefe und Briefe aus dem Briefroman kaum zu trennen sind, was insofern plausibel ist, als offenbar im Briefroman die Gattung des Privatbriefs literarisch imitiert wird. Fasst man also die Texte dieser beiden Textsorten in einer Gruppe zusammen und stellt sie den beiden anderen gegenüber (so dass man insgesamt drei Gruppen hat), erhöht sich der Prozentsatz korrekter Diskriminationen auf immerhin 86,90%. Eine weitere Zusammenfassung auch der beiden anderen Textsorten (offene Briefe und Leserbriefe) in einer gemeinsamen Gruppe führt dazu, dass nur noch zwei Gruppen übrigen bleiben, die sich bedingt als ‚private Briefe' vs. ‚öffentliche Briefe' bezeichnen lassen – in der Tat werden in dieser Form 92% aller Texte korrekt diskriminiert, was zu der Frage führt, inwiefern die so erhaltenen Obergruppen gegebenenfalls so etwas wie zwei Diskurstypen – einem privaten und einem öffentlichen – entsprechen. Sollte diese Vermutung zutreffen, sollten aber auch Texte der anderen Textsorten diesem Schema zuzuordnen sein.
Erweitert man folglich die Datenbasis im ersten Schritt um die Kommentare und ordnet diese dem Typ des öffentlichen Diskurses zu, so bleibt es auch bei den nunmehr 248 Texten bei 91,10% korrekter Diskriminationen. Die Wiedereinführung der dramatischen Texte erhöht die Datenbasis auf insgesamt 290 Texte; bei deren Zuordnung zum privaten Diskurstyp bleibt der Prozentsatz korrekter Diskriminationen mit 92,40% annähernd gleich hoch. Die Wiedereinführung auch der Gedichte als eigener Textsorte (Poesie) resultiert in einer in drei Gruppen untergliederten Gesamtmenge von 330 Texten. Auch unter dieser Bedingung bleibt der Prozentsatz korrekter Diskriminationen auf dem nach wie vor hohen Ni-

veau von 91,20%. Damit sind immerhin sieben der acht Textsorten berücksichtigt, und man kann sagen, dass die einzelnen Texte allein aufgrund ihrer Wortlänge mit mehr als 90% Trefferquote einem der drei Diskurstypen, wie wir sie genannt haben, zugeordnet werden können. Offen bleibt die Frage, wie es mit den literarischen Texten steht, die es als letztes wieder einzuführen gilt, was der Ausgangsbasis unserer 398 Texte entspricht. Führte man die 68 literarischen Prosatexte als eigene Gruppe wieder in die gesamte Textbasis ein (was einer Gesamtmenge von vier Diskurstypen entspräche), so resultierte das in einem vergleichsweise schlechten Ergebnis von 79,90% korrekter Diskriminationen.
Grund für diese Verschlechterung des Ergebnisses ist die Tatsache, dass die Mehrheit der literarischen Texte nicht einem eigenen literarischen, sondern dem privat-mündlichen Texttyp zugeordnet wird. Mögliche Gründe dafür könnten entweder darin zu sehen sein, dass sowohl die Privatbriefe als auch die literarischen Erzählungen im konkreten Fall von ein und denselben Autoren stammen, oder aber darin, dass es sich bei den ausgewählten literarischen Texten um solche handelt, die einen hohen Anteil an dialogischer Rede aufweisen und zudem in den narrativen Sequenzen einen mündlichen Redestil fingieren – hier werden nur weitere Untersuchungen Aufschluss bieten können.
Einstweilen liegt es nahe, die literarischen Texte der Gruppe der privat-mündlichen zuzuordnen, was es bei einer Unterteilung des gesamten Textmaterials in drei Obergruppen belässt – man erinnere sich, dass auch die Cluster-Analysen ja eine Unterteilung der gesamten Textbasis in drei Gruppen als optimale Clusterbildung nahe gelegt hatten. Im Gesamtergebnis führt dies jedenfalls dazu, dass sich von den 398 Texten allein aufgrund der Wortlänge 92,70% einem dem drei Diskurstypen korrekt zuordnen lassen (vgl. Abb. 4).

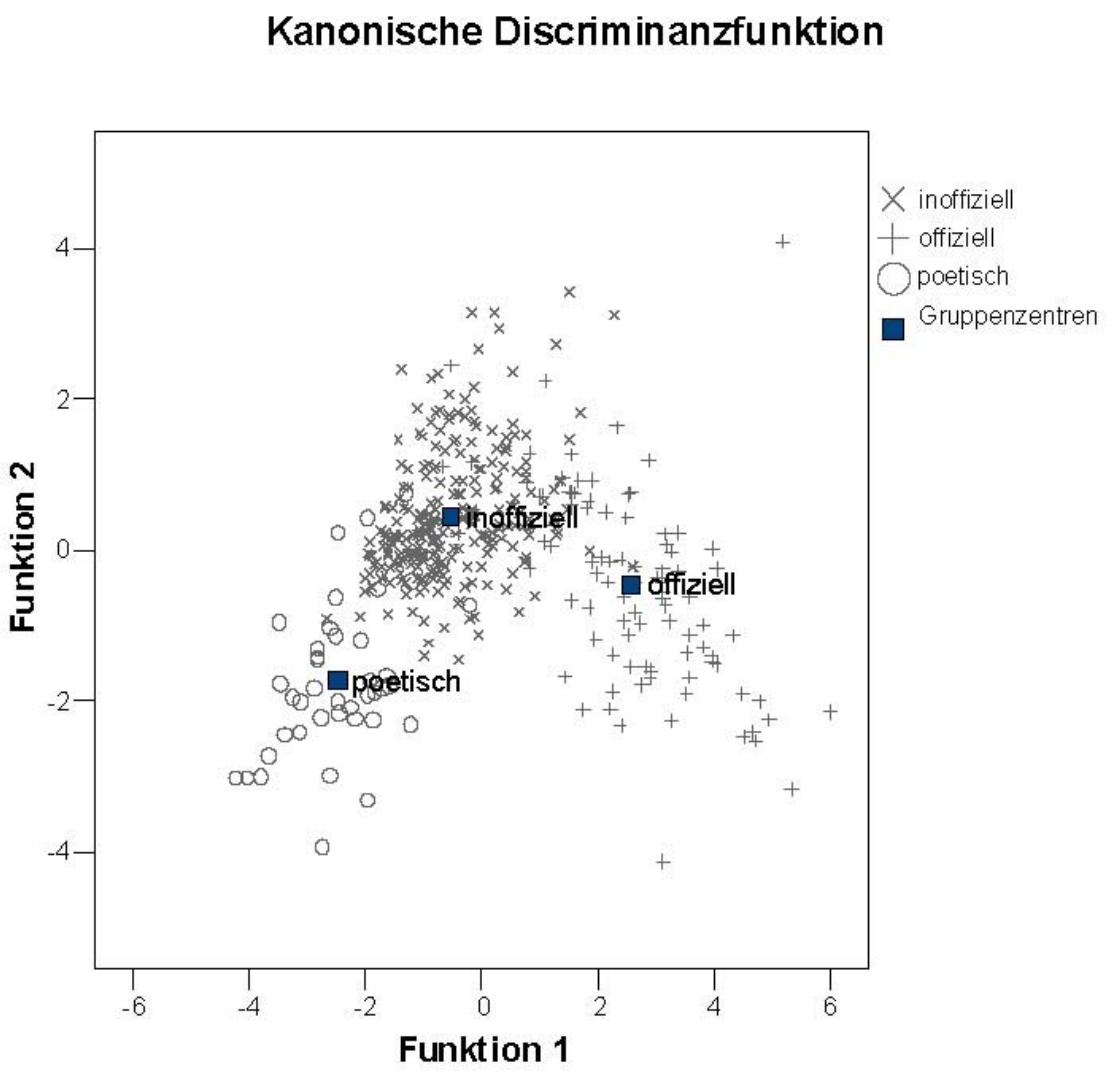


Abb. 4: Diskriminanz-Analyse (398 Texte)

In jedem Fall ist festzuhalten, dass die erhaltene Gliederung der von uns untersuchten Textwelt einzig und allein aufgrund der Wortlänge zu folgenden Diskurstypen führt:

a.) einem Privat- bzw. Alltagstyp,

b.) einem öffentlich/offiziellen Typ und

c.) dem poetischen Diskurstyp der Gedichte.

Dieses Ergebnis sollte jedoch nicht so verstanden werden, als dass möglicherweise drei Diskurstypen die Textwelt in einer optimalen Weise wiedergeben (was sowohl die Befunde der Cluster-Analysen als auch der multivariaten Diskriminanzanalysen nahe legen). Eine solche Vermutung kann erst durch die Analyse von weiteren Textsorten bestätigt wer-

den; dies ist insofern einstweilen als offene Fragestellung zu betrachten. In dieser Hinsicht angedeutet werden kann jedoch folgendes: Bei einer Erweiterung der Textbasis um 30 slowenische Kochrezepte – was nunmehr immerhin neun Textsorten in vier unterschiedlichen Diskurstypen beinhaltet – stellt sich heraus, dass bei Berücksichtigung der Textsorte in einem eigenen Diskurstyp ein unverändert hoher Anteil von 91,80% der Texte korrigiert diskriminiert werden (vgl. Grzybek/Kelih 2004a). Dies könnte gegebenenfalls auf einen eigenen Diskurstyp „Fachsprache“ hinweisen. Damit ist ein nicht gerade als schmal zu bezeichnender Spektralbereich des Textuniversums berücksichtigt; inwiefern weitere Textsorten dieses Gesamtbild ändern können, ist eine an weiteren Textsorten und vor allem auch weiteren Sprachen zu prüfende Frage.

Literatur

Altmann, G. (1992): Das Problem der Datenhomogenität. In: B. Rieger (Hrsg.), Glottometrika 13, 287-298.

Antić, G.; Kelih, E.; Grzybek, P. (2004): Zero-syllable Words in Determining Word Length. In: P. Grzybek (Ed.): Contributions to the Science of Language. Word Length Studies and Related Issues. New York. [In print]

Grzybek, P.; Kelih, E. (2004a): Texttypologie in/aus empirischer Sicht. In: J. Bernard; P. Grzybek, & Ju. Fikfak (Eds.): Text and Reality. Ljubljana etc. [In print].

Grzybek, P.; Kelih, E. (2004b): Wortlänge in Silben, Graphemen, Morphemen (am Beispiel russischer Texte). In: Grzybek, P. / Kelih, E. (Eds): Wörter, Längen, Häufigkeiten. [In Vorb.]

Grzybek, P.; Stadlober, E. (2003): Zur Prosa Karel Čapeks – Einige quantitative Bemerkungen. In: S. Kempgen, U. Schweier, T. Berger (Eds.), Rusistika – Slavistika – Lingvistika. Festschrift für Werner Lehfeldt zum 60. Geburtstag. Sagner, München, 474-488.

Grzybek, P.; Stadlober, E.; Kelih, E.; Antić, G. (2004): Quantitative Text Typology: The Impact of Word Length. In: Weihs, C. (ed.): Classification – The Ubiquitous Challenge. Heidelberg/Berlin. [In print]

Kelih, E.; Antić, G.; Grzybek, P., Stadlober, E. (2004): Classification of Author and(or) Text? In: Weihs, C. (ed.): Classification - The Ubiquitous Challenge. Heidelberg/Berlin. [In print]

Köhler, R. (1986): Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. Studienverlag Brockmeyer, Bochum. (= Quantitative Linguistics, 31)

Köhler, R. (1998): Vorwort. In: Tuldava, J.: Probleme und Methoden der quantitativ-systemischen Lexikologie: i-iv. Trier.

Mistrík, J. (1973): Eine exakte Typologie von Texten. (= Arbeiten und Texte zur Slavistik 3), München.

Orlov, Ju. K. (1982). Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? Die Antinomie "Sprache – Rede" in der statistischen Linguistik. In: Orlov, Ju. K., Boroda, M. G., Nadarešvili, I. Š. (1982): Sprache, Text, Kunst. Quantitative Analysen: 1-55. Bochum.