BMBF

Verbmobil

Verbundvorhaben

Klassifikation von Sprechstilen mit linguistischem Wissen

Martin Klarner

Friedrich-Alexander-Universität
Erlangen-Nürnberg

Vm

REPORT 214
Februar 1998

Februar 1998

Martin Klarner
Institut für Mathematischen Maschinen und Datenverarbeitung VIII
Lehrstuhl für Künstliche Intelligenz
Friedrich-Alexander-Universität Erlangen-Nürnberg
Am Weichselgarten 9
91058 Erlangen
Tel.: (09131) 8599 - 13
Fax: (09131) 8599 - 05
e-mail: `klarner@immd8.informatik.uni-erlangen.de`

Gehört zum Antragsabschnitt: 2.2.6 Sprechstiladaption

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 G gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.
ISSN 0949-6084

Zusammenfassung

Die Unterscheidung zwischen Inhalt und Form läßt sich insbesondere auch für die Sprachproduktion treffen; die sprecherspezifische Ausprägung der Form von Äußerungen bezeichnet man als persönlichen Sprechstil. Im Unterschied zur Stilanalyse schriftlicher, speziell literarischer Texte, wo in der Regel ausreichend Material eines Autors vorhanden ist, ist die zur Verfügung stehende Textmenge pro Sprecher bei der Sprechstilanalyse oft sehr gering. In dieser Arbeit wird versucht, individuelle Sprechstile mit den Mitteln der quantitativen Stilistik zu identifizieren, mit Methoden der Clusteranalyse zu klassifizieren und das daraus gewonnene Wissen für Sprachverarbeitungssysteme nutzbar zu machen.

Ergebnis der Untersuchungen sollen Merkmale sein, mit denen sich ein unbekannter Sprecher einer Klasse gleichartiger Sprecher zuordnen und das Sprachverarbeitungssystem an diesen Sprecher anpassen läßt.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

1.1 Allgemeine und quantitative Stilistik

Mit dem Begriff Sprachstil und den damit zusammenhängenden Termini „stilistische Besonderheiten“, „Stilmerkmale“ u. a. verhält es sich wie mit jedem alltagssprachlichen Wort, das Eingang in die Wissenschaft gefunden hat: Die Ansätze zu einer exakten Definition sind ebenso vielfältig wie heterogen[1].

Beim Stil ist diese Situation vielleicht wirklich unvermeidlich, weil Sprache und Stil Untersuchungsgegenstand so unterschiedlicher Fachdisziplinen wie Literaturwissenschaft und Computerlinguistik, Informatik (hier insbesondere die automatische Sprachverarbeitung in KI und Mustererkennung) und allgemeiner Sprachwissenschaft, ja sogar Psychologie und Philosophie sind. Befassen wir uns aber nicht mit einer tiefergehenden Erörterung dieser Thematik, sondern verweisen lieber auf die einschlägige Literatur, z.B. [44] und [32]. Hier soll der Hinweis auf die beiden Hauptströmungen genügen, die SANDERS den „selektiven“ und den „statistischem Stilbegriff“ nennt (vgl. [32], S. 17 ff.). Die zur ersten Richtung zu rechnenden Definitionen betonen den Gedanken der Auswahl unter konkurrierenden, „informationsadäquaten“[2] Sprachmitteln unter dem Gesichtspunkt der „Zweckmäßigkeit“[3]. Augrund des selektiven Charakters stilistischer Besonderheiten vergleicht WINTER in [43] Stile mit Soziolekten, die im Unterschied zu Dialekten aus dem Formenbestand der gesamten Sprachgemeinschaft schöpfen können. Neben dem Auswahlprinzip wird Stil andererseits auch vielfach durch die Abweichung von einer „Norm“ definiert, wobei diese aber schwer festzulegen ist (vgl. [15], S. 210).

Während sich die Literaturwissenschaft naturgemäß mit stilistischen Erscheinungen in schriftlich vorliegenden Texten[4] bestimmter Autoren oder Epochen, Schichten oder Gruppen[5] befaßt, wird in [32], S. 38 ff. der Stil *gesprochener* Sprache im Kontrast zum schriftsprachlichen Stil gesehen[6]. SANDERS betont hier den Primat der gesprochenen Sprache vor der Schriftspra-

[1] Für eine Zusammenstellung siehe z.B. [32], S.17 ff., für einen historischen Überblick z.B. [15], S. 195 ff.

[2] dies die Sicht- und Ausdrucksweise des Autors dieser Arbeit, der seinen informatischen Hintergrund nicht verleugnet

[3] auch dieser Begriff wird von Fall zu Fall unterschiedlich definiert

[4] Beim Textbegriff ist im übrigen in gleicher Weise zu beobachten, wie hoch der Grad der Verschriftung unserer Kultur ist, so daß eine Wendung wie „schriftliche Texte“ schon als *contradictio in adiecto* angesehen wird; gleichwohl gilt auch hier, daß Texte durchaus nicht in geschriebener Form vorliegen müssen: Die Textlinguistik hat bei der Definition ihres Forschungsgegenstandes mit den gleichen Problemen zu kämpfen wie die Stilistik

[5] an sprachlich-stilistischen Einzeluntersuchungen von Philologen zu klassischen wie modernen Autoren, zu historischen Epochen wie zu unterschiedlichen Regionen und Kulturen herrscht kein Mangel

[6] Aus der Sicht des Sprachverarbeiters wäre hinzuzufügen, daß mit „gesprochener Sprache“ spontansprach-

che und sieht die typischen Merkmale der ersteren, also Rekurrenzen verschiedener Art, Anakoluthe, Ellipsen u. a., nicht etwa als Fehler an, wie dies explizit in der traditionellen Grammatiktheorie und in subtilerer Form in der Transformationsgrammatik geschieht[7]. Stilistische Merkmale finden sich beim einzelnen Sprecher in individueller Weise, die den persönlichen Sprechstil ausmacht — jener bei vielen Autoren und Rednern ebenso deutlich erkennbaren wie schwer zu beschreibenden und objektivierbaren Eigenart des Sprachgebrauchs, die nicht in den Gegebenheiten der Kommunikationssituation und in dem Wunsch nach unmißverständlicher Informationsübermittlung begründet ist, sondern ihre Ursache in der freien Wahl der Sprachmittel und im je eigenen Umgang mit der Sprache hat.

Der grundsätzlich andere Ansatz des statistischen Stilbegriffs beruht auf der Auffassung begründet, daß Stil stets als eine *Kombination von quantitativ erfaßbaren Merkmalen* gedeutet werden kann. Die Linguistik gewinnt an dieser Auffassung eine objektive und nachvollziehbare Methodik zur Stilbeschreibung, die Literaturwissenschaft den Übergang von der bloßen Benennung und Beschreibung von Begriffen zur Erklärung. Man kann die Situation in der Sprachwissenschaft als einer typischen geisteswissenschaftlichen Fachdisziplin folgendermaßen zusammenfassen:

> „Qualitative Begriffe erscheinen zunächst natürlicher zu sein [...] In fortgeschritteneren Stadien einer Wissenschaft kommt man aber ohne Quantifizierung nicht aus. [...]“[8]

Es kann nicht überraschen, daß die statistische Stilistik von Literaturwissenschaftlern skeptisch beurteilt[9] und die sicherliche etwas exzentrische Auffassung des statistischen Stilbegriffs als der „adäquaten Theorie des Stils“ natürlich rundweg abgelehnt wird. Wir müssen an dieser Stelle nicht so weit gehen, weil nicht etwa die — wie auch immer zu definierende — Adäquatheit statistischer Stiltheorien Gegenstand dieser Arbeit ist, sondern ihre Anwendbarkeit in und ihr Nutzen für die Erkennung gesprochener Sprache[10]. Festzuhalten bleibt jedoch, daß die im folgenden angenommene Stildefinition eben die der quantitativen Stilistik in ihrer „reinen Form“ zu sein hat, und zwar schon aus praktischen Überlegungen, nicht nur der objektiven Nachvollziehbarkeit, sondern auch der Einsetzbarkeit in konkreten Systemen.

1.2 Resultate aus angrenzenden Bereichen

Die literaturwissenschaftliche Beschäftigung mit Stil zeigt sich exemplarisch in [36] und in [33]: Es werden einerseits die klassischen „mikrostilistischen“ Elemente, im wesentlichen die traditionellen Stilfiguren und Tropen, und andererseits „makrostilistische“ und „pragmatische“ Eigenschaften betrachtet, also satzübergreifende, zum Teil den gesamten Text umfassende

liche Äußerungen gemeint sind; es existiert ja auch eine Reihe von Formen nichtspontaner und dennoch nichtschriftlicher Sprache, wie Reden, Lesungen und Vorlesungen, Monologe etc.

[7] Im übrigen liegt die hauptsächliche Motivation für den Einsatz moderner stochastischer Grammatiken in der Sprachverarbeitung gerade in der Erkenntnis, daß spontansprachliche Äußerungen aufgrund ihrer Irregularitäten und Abweichungen von der Schriftsprache eben nicht von einer klassischen Grammatik mit fester Regelstruktur adäquat verarbeitet werden können.

[8] Altmann in [3], S. 113

[9] Ein typisches Beispiel findet sich in [44], S. 30. Man hat zuweilen den Eindruck, Literaturwissenschaftler fürchteten, sie würden sich durch eine Akzeptanz der Statistik selbst überflüssig machen; natürlich könnte nichts der Wahrheit ferner liegen.

[10] Wir werden dieses umfassende Ziel gleich auf den Rahmen einer Diplomarbeit reduzieren müssen.

Phänomene. Während der zweite Bereich einer quantitativen Analyse weitgehend unzugänglich ist, läßt sich der erste auf geschriebene Texte und auch auf vorformulierte und -konzipierte Reden anwenden; in der Alltagssprache werden klassische Figuren wohl seltener anzutreffen sein.

Zweifellos gute Dienste leistet die statistische Stilistik bei der Beantwortung strittiger Fragen in klar abzugrenzenden Bereichen wie Taxonomien von Textsorten und Autorenschaftsuntersuchungen. Man findet hier in der Regel die für quantitative Fragestellungen ideale Situation vor: Es sind mehr als genügend gesicherte Daten der Textgruppe bzw. des Autors in Form eindeutig ihr zuzuordnender bzw. von ihm verfaßter Texte verfügbar, anhand derer sich die charakteristischen Merkmale des Textgruppen- bzw. Autorenstils eindeutig ermitteln lassen. Meist gibt auch die Sprachwissenschaft ihren Segen in Form einer treffenden linguistischen oder literaturwissenschaftlichen Interpretation des statistischen Materials, so daß für einen untersuchten Text ohne weiteres eine passende Textsorte angegeben bzw. er einem Autor zugeordnet oder als nicht von ihm verfaßt erkannt werden kann. Natürlich kann es auch hier Problemfälle geben, wie z.B. klassische Textfragmente, bei denen die in Frage kommenden Autoren nicht genügend Brauchbares[11] hinterlassen haben, aber auch dann kann man sich mit entsprechend robusten statistischen Verfahren meist noch behelfen und zu guten Resultaten kommen.

Ganz anders als bei den üblicherweise mit den Mitteln der quantitativen Stilistik untersuchten Fragestellungen liegen die Dinge bei dieser Arbeit, die sich von folgenden „Arbeitshypothesen“ her motivieren läßt:

> *Hypothese I:* Sprecher lassen sich aufgrund des in ihren spontansprachlichen Äußerungen manifesten persönlichen Stils in Klassen einteilen.
>
> *Hypothese II:* Diese Klasseneinteilung kann durch quantitativ-stilistische, linguistisch fundierte Merkmalssätze ermittelt werden.
>
> *Hypothese III:* Damit lassen sich an den Sprecherstil adaptierte und spezialisierte Sprachmodelle bestimmen, wodurch sich die Leistung von Sprachverarbeitungssystemen verbessert.

Insbesondere zwei gravierende Unterschiede zur vorher geschilderten Situation fallen ins Auge:

1. Aus der Sicht der Linguistik betrachten wir in dieser Arbeit nicht geschriebene Texte, nicht einmal in irgendeiner Weise vorgefaßte und vielleicht stilistisch überarbeitete Äußerungen, sondern *spontane Sprache*.

2. Der Informatiker, insbesondere der mit den Problemen der Mustererkennung vertraute Leser wird sofort sehen, daß wir nicht über die für ein „überwachtes Lernen“ entscheidende klassifizierte Stichprobe verfügen: Es gibt keine vorgefertigte Trainingsmenge für einen Klassifikator mit geeigneten Sprecherklassen, weil nicht einmal gesichert ist, daß eine objektive, d.h. auf möglichst alle Sprecher anwendbare und auch linguistisch motivierte Klasseneinteilung existiert, geschweige denn wie viele Klassen sie umfaßt und wie diese zu charakterisieren sind[12].

[11] in statistischer Hinsicht, wohlgemerkt, d.h. Fragmente mit teilweise fragwürdiger Autorenschaft

[12] Bereits an dieser Stelle wird deutlich, daß — mit einigem Understatement — die Auswahl geeigneter Merkmale ein Problem darstellen könnte.

Der erste Punkt erzwingt eine Abkehr von jeglicher literaturwissenschaftlichen Stildefinition und die Hinwendung zur anderen Hauptrichtung, der statistischen Stiltheorie, der zweite die Benutzung von Verfahren des *unüberwachten Lernens*, insbesondere der Clusteranalyse einer aus den verfügbaren Daten zusammengestellten Trainingsmenge mit Hilfe noch zu definierender Merkmale.

Daneben läßt sich noch konstatieren, daß der Linguist, wenn er sich mit quantitativer Stilistik befaßt, i. a. nicht an konkreten Einsatzmöglichkeiten in einem Sprachverarbeitungssytem interessiert ist, sondern für ihn der deskriptive Aspekt seiner Sprachmodellierung im Vordergrund steht[13].

Wenn auch die Situation dieser Arbeit anders als die üblicherweise für quantitative Stiluntersuchungen gegebene ist, so wurden doch verschiedene Teilaspekte der in den Arbeitshypothesen angesprochenen Probleme bereits untersucht; hier eine Auswahl:

In [14] wird die Möglichkeit einer Stilanalyse „interaktiven spontanen Sprachgebrauchs"[14] betrachtet. Die Autorin weist auf die Rolle der funktionalen Stilistik hin, die im Spannungsfeld zwischen der literarischen Stilauffassung, daß in einem Text alle Elemente bewußt gesetzt sind, und der psychologischen Ansicht, daß Stil dem Sprecher unbewußt „unterläuft", die situativen und zweckgerichteten Aspekte stilistischer Merkmale in den Vordergrund stellt.

Franck sieht Stil als „Resultat einer bestimmten Betrachtungsweise des Textes" ([14], S. 122) und betont die pragmatische Funktion stilistischer Elemente, die unter den Produktionsbedingungen sprachlicher Äußerungen zu sehen sind: Stil wird als Auswahl aus mehreren inhaltlich gleichen Möglichkeiten im Hinblick auf den Adressaten und die Kommunikationssituation verstanden (*recipient design*), der durch „implizites Aushandeln" determiniert wird. Dies geht bis zum gegenseitigen Aufgreifen von Formulierungen und Stilfiguren trotz adäquaterer Ausdrucksalternativen (vgl. [14], S. 134). — Hier stellt sich natürlich sofort die Frage, ob unter diesen Umständen ein unabhängiger persönlicher Sprecherstil überhaupt noch angenommen oder ob spontansprachlicher Stil nicht vielmehr ausschließlich im Dialog untersucht und beschrieben werden kann[15].

Zwar werden von Pieper in [30] nicht explizit stilistische Untersuchungen betrieben, jedoch soll ihre Arbeit hier als Beispiel für den Ansatz dienen, mit multivariater Statistik zur Charakterisierung und Klassifikation von Textgruppen zu gelangen; dabei interesieren hier mehr die verwendeten Merkmale und Methoden als die Ergebnisse der Textgruppendifferenzierung, weil letztere nicht auf die Thematik dieser Arbeit übertragbar sind; erstere jedoch werden im weiteren Verlauf der Untersuchungen noch einige Male begegnen.

Die Autorin verwendet Merkmale aus Morphologie und Syntax, die sämtlich auf den Satz, der für Texte charakteristischen Einheit, bezogen sind. Sie gelangt dabei von der Betrachtung einzelner Merkmale über Quotienten als „klassischem" Ansatz zur Berücksichtigung mehrerer Variablen bis zur Darstellung multivariater Zusammenhänge durch Regressionsanalysen und Kausalmodelle mit Hilfe von Pfadanalysen[16]. — Auch Häufigkeitsverteilungen der

[13] und er sich dabei also auch nicht der quantitativen Bewertung seiner Resultate zu stellen hat, wie das für die Ergebnisse dieser Arbeit in 4.3 geschehen wird

[14] [14], S. 121

[15] Die erste Alternative wollen wir in dieser Arbeit betrachten, zur Untersuchung der zweiten ist, wie wir unten noch sehen werden, das verfügbare Datenmaterial nicht ausreichend.

[16] So weit werden wir in dieser Arbeit nicht gehen, weil hier weniger die funktionalen Zusammenhänge zwischen stilistischen Charakteristika von Bedeutung sind als das Auffinden trennstarker Merkmale zur Klassifikation.

Hauptkategoriekomplexe Nomen, Verb und Adjektiv werden betrachtet, um die untersuchten Textgruppen grob- und feindifferenziert zu klassifizieren.

Bei der Begründung der linguistischen Relevanz der erarbeiteten Befunde verwahrt sich die Autorin gegenüber literaturwissenschaftlichen Einwänden gegen die quantitative Methodik, indem sie eine „Synthese beider Richtungen“ ([30], S. 125) anstrebt, also über die bloße Auflistung von statistischen Resultaten hinaus eine linguistische Interpretation der Ergebnisse anstrebt: Dies wird auch ein Anliegen der vorliegenden Arbeit sein.

Lexikalische Merkmale stehen in [7] bei einer Textgruppendifferenzierung zwischen den Gattungen der englischen Sprache „Trivialliteratur“ und „literarisch anspruchsvoller Roman“ im Vordergrund: BOLZ kommt zu dem Schluß, daß sich mit Wortlängenmerkmalen die beiden zu unterscheidenden Gattungen am besten voneinander trennen lassen. Dies kann selbstverständlich eine Eigenheit des Englischen oder der betrachteten Texte sein, aber man tut gut daran, Wortlängenmerkmale auch im Kontext der vorliegenden Arbeit zu berücksichtigen — schon weil sich einfach und relativ rasch aussagekräftige Werte ermitteln lassen.

Der Aspekt einer Aufteilung der vorhandenen Daten in Subkorpora mit dem Ziel, die Leistung eines Sprachverarbeitungssystems zu verbessern, wird bereits in [31] für ein natürlichsprachliches Dialog-Auskunftssystem untersucht. Allerdings wird die Klassifikation der Trainingsdaten nicht mit quantitativen Verfahren ermittelt, sondern auf eine „natürliche Weise“ vorgegeben. Diese ist auch nicht stilistisch motiviert, sondern berücksichtigt die Dialogakttypen von Fragen des Systems an den Benutzer. Dennoch kann der Ansatz, mehrere spezielle Sprachmodelle statt eines globalen Modells zu verwenden und dabei den Balanceakt zwischen Spezialisierung und Robustheit[17] erfolgreich zu vollziehen, als eine Motivation dieser Arbeit gelten.

In [4] dagegen wird die sprecherspezifische Anpassung eines einzelnen Sprachmodells in einem Diktiersystem beschrieben, das mit einem globalen Modell verknüpft wird, um auch mit relativ wenigen sprecherspezifischen Trainingsdaten bereits gute Ergebnisse bei der Spracherkennung zu erzielen[18].

Trotz aller von anderer Seite geleisteten Vorarbeiten sind einschlägige quantitativ-stilistische Untersuchungen zum sprecherspezifischen spontansprachlichen Stil in der Literatur nicht auffindbar; darin liegt jedoch gerade eine Motivation zu dieser Arbeit, deren zentrale Fragestellung wir hier noch einmal formulieren wollen:

Gibt es Klassen persönlicher Sprachstile, und lassen sie sich adäquat, d.h. quantitativ signifikant und linguistisch nachvollziehbar, beschreiben?

[17] Wir werden auf dieses sog. Parameterproblem in 4.1 noch weiter eingehen.
[18] Auch auf diesen Ansatz werden wir gegen Ende unserer Ausführungen noch einmal zurückkommen.

Kapitel 2

Von Variablen zu Merkmalssätzen

2.1 Das VERBMOBIL-Korpus

Beim VERBMOBIL-Korpus (im folgenden mit VM-Korpus abgekürzt), dessen sich die vorliegende Arbeit als Textgrundlage bedient, handelt es sich um eine Sammlung spontansprachlicher Dialoge, die über Mikrofone in getrennten Räumen geführt wurden (vgl. [24]). Die Sprecher waren dazu aufgerufen, in den Dialogen mit Hilfe eines vorgegebenen Terminkalenders zu einer Absprache über gemeinsam abzuhaltende Veranstaltungen, geschäftliche Treffen u. ä. zu kommen. — Das Korpus umfaßt Audiodaten von 623 Sprechern, verteilt über 7 CDs, mit insgesamt 12046 Turns, 26236 Sätzen und 253357 Wörtern.

Die daraus ausgewählte Trainingsstichprobe, auf der die Tests zur Ermittlung von Merkmalen, die spätere Klassifikation und auch die Erstellung von Sprachmodellen und die Bewertung durchgeführt wird, umfaßt die CDs 1, 3, 5, 7 und 12 des VM-Korpus. Sie enthält 9018 Turns, 19760 Sätze und 192843 Wörter von 470 Sprechern.

Die Dialoge wurden nach den orthographischen Regeln unter Einschluß der Interpunktion transliteriert. In diese Transliterationen wurden auch nonverbale Produktionen der Sprecher, wie Häsitationen, Lachen, Redepausen etc. mit aufgenommen. Ebenfalls enthalten sind Satzabbrüche und Wiederaufnahmen, regionale Aussprachevariantenund Neologismen, Überlappungen von Turns und akustische Nebengeräusche (vgl. [24], S. 33), so daß sich aus der Transliteration ein recht umfassendes Bild des Dialogverlaufs über die reine Wortfolge hinaus ergibt. Die Dialogdaten wurden außerdem mit POS-Tags nach den STTS-*guidelines* (vgl. [34]) versehen.

Zusätzlich wurden die Transliterationen von Hand mit pragmatischen Dialogakten gemäß [20] annotiert und syntaktische Bäume der einzelnen Sätze (sog. *chunktrees*) erstellt.

Für diese Arbeit standen also an Daten zur Verfügung bzw. wurden aus den Transliterationen erzeugt:

- die vollständig transliterierten Dialoge
- gefilterte Transkriptionen, die keinerlei nonverbale Produktionen[1] mehr enthalten
- die mit POS-Tags annotierten Dialoge
- mit Hauptkategorien[2] annotierte Dialoge

[1] bis auf Häsitationen

[2] die sich aus einer Zusammenfassung der POS-Tags ergeben

- Annotationen von Dialogen mit Dialogakten
- syntaktische Bäume von Sätzen

Das VM-Korpus weist als Kollektion spontansprachlicher Dialoge gegenüber den meisten Textkorpora, die in den eingangs zitierten Arbeiten untersucht wurden, einige Besonderheiten auf, die zum Teil unmittelbar einsichtig sind, zum Teil auch als Erwartungen formuliert werden müssen, die das Vorgehen in eine bestimmte Richtung lenken, aber zur Gewinnung konkreter Aussagen genauer überprüft werden. Sofern diese Besonderheiten für die weiteren Untersuchungen von Belang sind, sollen sie im folgenden skizzenhaft und weitgehend qualitativ aufgeführt werden.

2.1.1 Spezifische Eigenschaften

Vokabular und Syntax

Aufgrund der Domäne des VM-Korpus finden sich sehr häufig Zeit-, Datums- und Ortsangaben; dies ist intuitiv klar und bedarf keiner weiteren Untersuchung.

Weil die spontansprachlichen Äußerungen in Dialogform erfolgen, ist eine hohe Frequenz von Pronomina zu erwarten, die auf Kosten der Substantive geht: In einer Rangliste der zehn häufigsten Vokabeln nehmen die Pronomina „ich“ und „mir“ vordere Plätze ein, das häufigste Substantiv findet sich mit „Uhr“ erst deutlich weiter unten auf Rang 22.

Daß der Vokabularumfang spontansprachlicher Äußerungen deutlich unter dem schriftlicher, geschweige denn literarischer Texte liegt, bedarf keiner ausführlichen Begründung. Zwar lag dem Autor keine Lemmatisierung des VM-Korpus vor, doch die Zahl von 5697 gefundenen Wortformen, von denen die 634 Eigennamen[3] noch zu subtrahieren sind, bei einer Gesamtwortzahl von über 192000, stützt diese Annahme.

Was die Modi anbelangt, so sind nicht allzuviele Imperative zu erwarten: Aufforderungen, so sie in Verhandlungsdialogen vorkommen, erfolgen in der Regel im höflichen Konjunktiv, während ansonsten der Indikativ vorherrscht.

Auch die Tempora sind in spontanem Sprachgebrauch weit weniger differenziert als in schriftlichen Texten: Das Präsens spielt die dominierende Rolle; daneben ist nur noch das Perfekt in angemessener Zahl zu erwarten, weil Futurformen umgangssprachlich meist durch präsentische ersetzt werden und Präterita ohnhehin selten anzutreffen sind.

Die Sätze sind länger als bei gesprochener Sprache üblich; der von MEIER in [26], S. 191 für Dialoge als Standard ermittelte 5-Wort-Satz wird deutlich überschritten, die Satzlänge des VM-Korpus liegt im Mittel bei knapp 10 Wörtern. Trotzdem finden sich, wie für Spontansprache üblich, überwiegend parataktische Konstruktionen.

Normalerweise erwartet man in Dialogen, auch in relativ formell geführten Verhandlungsdialogen, von Zeit zu Zeit eine Unterbrechung durch den Gesprächspartner. Im VM-Korpus fällt laut Transliteration jedes Turnende mit einem Satzende zusammen, m. a. W. es gibt de facto keine Unterbrechungen. Allerdings finden sich Überlappungen von Turns, die aber nicht als Unterbrechungen interpretiert werden können, da ja die Sätze trotzdem zu Ende geführt werden[4].

[3] die zu Testzwecken für die Spracherkennung bewußt evoziert wurden, vgl. [24], S. 5)

[4] Dieses Fehlen von Unterbrechungen läßt sich auch mit der Gesprächs- und Aufnahmesituation begründen: Bei der von vornherein auf Kooperation angelegten, ja angewiesenen Domäne der Terminabsprache läßt man den Gesprächspartner eben ausreden.

Statistik

Aus statistischer Sicht ist das vorhandene Datenmaterial des VM-Korpus, so umfangreich es auf den ersten Blick sein mag, im Hinblick auf die einzelnen Sprecher in vielen Fällen zu dürftig. Wenn man die Situation mit der bei einer Autorenschaftsuntersuchung oder gar einer Textgruppendifferenzierung[5] vorliegenden vergleicht, so stellt sich bei jedem der weiter unten erläuterten Merkmale die Frage, ob die Bedingungen für ein aussagekräftiges Resultat noch erfüllt sind.

Wir werden deshalb bei der Untersuchung der Merkmale noch näher auf diese Problematik eingehen; hier seien lediglich zwei Beobachtungen zur Illustration angemerkt: Zum einen ergibt sich nmittelbar aus der Wortzahl von gut 192000 und der Sprecherzahl von 623, daß die Wortzahl pro Sprecher im Mittel nur bei etwas über 300 WÖrtern liegt. — Zum anderen reicht der Umfang der verfügbaren Daten der einzelnen Sprecher von 5 bis zu über 100 Turns. Jedoch kann man es sich kaum leisten, auf Sprecher mit zu wenigen Turns völlig zu verzichten, weil beispielsweise eine ausschließliche Berücksichtigung von Sprechern mit mindestens 20 Turns zum Verlust von 2/3 der verfügbaren Sprecher führen würde.

2.1.2 Implikationen für die Untersuchung

Eine exakte Darstellung der Besonderheiten des VM-Korpus ist jedoch nicht Gegenstand der vorliegenden Arbeit; daher wurden auch, wie angekündigt, lediglich diejenigen Spezifika angesprochen, die von Bedeutung für die nachfolgenden Untersuchungen sind.

Grundsätzlich läßt sich sagen, daß quantitative Restriktionen und qualitative Spezifika die Verwendung von bestimmten Merkmalen ausschließen, obwohl sie bereits von anderer Seite untersucht wurden und dabei für schriftliche Texte gute Ergebnisse erzielt haben. Anders formuliert, es lassen sich rasch a priori ungeeignete Merkmale und Merkmalsgruppen für eine Klassifikation der Sprecher des VM-Korpus ermitteln, was in diesem Abschnitt erfolgen soll.

Makrostilistik und Pragmatik: Dialogakte

Die Verwendung des aus linguistischer Sicht äußerst interessanten[6] makrostilistischen Merkmalskomplexes der Dialogakte, die im VM-Korpus aufgrund von Hand vorgenommener Annotationen zur Verfügung stehen, scheitert an der zu geringen Gesamtmenge des vorhandenen Materials: Längst nicht alle Dialoge sind annotiert, und selbst die vollständig annotierten Dialoge weisen, auf die einzelnen Sprechern verteilt, zu niedrige Werte auf.

Zwar ist die Gesamtmenge der Dialogakte hoch genug, aber sie ist so stark mit der Satz- bzw. Turnzahl korreliert, daß diese einfacher zu ermittelnden und vollständig verfügbaren Merkmale besser geeignet sind. Eine Betrachtung der einzelnen Typen von Dialogakten verbietet sich wegen ihrer zu geringen Anzahl sofort, aber sogar eine Gruppierung nach inhaltlichen Kriterien (etwa in orts- und zeitbezogene Dialogakte) ließ sich aus dem gleichen Grund nicht sinnvoll verwenden.

Wir müssen diese Aussagen auf den Bereich der Makrostilistik ausweiten, weil die Situation dort noch wesentlich unbefriedigender ist: Literarische, schwer quantifizierbare Eigenschaften wie Komposition, beschreibende vs. dialogische Darstellungsweise und Perspektive[7] sind einer quantitativen Betrachtung per definitionem unzugänglich und für Spontansprache ohnehin

[5] bei der die Datenmenge sogar potentiell unendlich ist
[6] nämlich aufgrund der Einbeziehung des Bereichs „Pragmatik“
[7] vgl. [36]; die Unbrauchbarkeit makrostilistischer Eigenschaften wurde ja bereits in 1.2 angedeutet

irrelevant. Sofern aber makrostilistische Merkmale quantitativ erfaßt werden können, tritt wieder das gleiche Problem wie bei den Dialogakten auf, daß im VM-Korpus zu wenige Daten pro Sprecher für eine statistische Untersuchung verfügbar sind.

Lexik

Einerseits gilt, „daß sich stilistische Variationen auf lexikalischer Ebene besonders in den Vokabeln mit schwacher Häufigkeit ausdrücken" ([29], S. 113), andererseits wird aber beispielsweise für einen aussagekräftigen χ^2-Test[8] als eine Faustregel eine theoretische Gesamtmenge von mindestens 5, besser 10 pro Einheit ([29], S. 119) angenommen. Selbst wenn man als Einheit die gesamte Datenmenge eines Sprechers ansieht, ist bei dem vorliegenden Datenmaterial die Frage, ob einzelne Vokabeln eine Untersuchung wert sind, eindeutig negativ zu beantworten. Sie wurden lediglich in der Zusammenfassung der „Füllwörter" (s. u.) verwendet.

Abstrakte Nomina und Fremdwortsubstantive, die durch ihre Endungen[9] kenntlich sind, machen bei Pieper 6 von 26 ausgewerteten Merkmalen aus (vgl. [30], S. 45). Wie aber bei spontaner Sprache wohl nicht anders zu erwarten, selbst von akademisch gebildeten Sprechern in einer offiziellen, geschäftlichen Gesprächssituation, finden sich diese Abstrakta derartig selten (insgesamt nur etwa 150 Vorkommen bei den 623 Sprechern), daß statistische Untersuchungen nicht lohnen.

Auch die Neologismen teilen dieses Schicksal: Obwohl es durchaus denkbar wäre, daß eine Sprecherklasse von einer stilistischen Grundhaltung geprägt ist, die man als „kreativen Umgang mit der Sprache" beschreiben könnte, lassen sich für eine vernünftige statistische Betrachtung und gar für eine Klassifikation bei weitem zu wenige Vorkommen nachweisen, nämlich mit einer Gesamtzahl von 187 im Schnitt 0,3 pro Sprecher. Müssen die Domäne und die Gesprächssituation wiederum als Erklärung für inhaltliche und formale Restriktionen bei der Sprachverwendung herhalten? Es scheint fast keine andere Erklärung denkbar, denn in einer geschäftlichen Terminverhandlung sind Neologismen salopp (d.h. entstammen einem unter dem situationsadäquaten Sprachregister angesiedelten Niveau) und verständnisreduzierend.

Spontansprachliche Einzelmerkmale

In der VM-Transkription sind eine ganze Reihe von Merkmalen gesprochener Sprache aufgeführt, die aus linguistisch-stilistischer Sicht von Interesse wären.

Die geringe Frequenz[10] vor allem im Verhältnis zu den Basiseinheiten Satz und Turn läßt eine Berücksichtigung der Einzelwerte nicht zu, wohl aber eine solche in Zusammenfassungen, die möglichst linguistisch motiviert vorgenommen wurden; näheres hierzu bei der Merkmalsauswahl in 2.2.3.

Verbale Flexionsformen: Tempus, Modus, Diathese

Die eindeutige Bestimmung von Tempus und Modus ist mit Ausnahme des Perfekts (indirekt über die Partizipialtags) und des Imperativs mit den Tags des STTS-Sets nicht möglich; gleiches gilt für die Diathese.

[8] Wir werden auf die in dieser Arbeit verwendeten statistischen Tests in 2.2.2 zu sprechen kommen.

[9] „-keit", „-heit", „ismus", „-ion", "-ik" und „-ung" werden in [30] untersucht

[10] z.B. von Lachen; einerseits bedauerlich, andererseits wiederum durch die Domäne und die vorgegebene wie die tatsächliche Gesprächssituation erklärbar

Während letztere für Spontansprache nicht weiter von Bedeutung ist, weil Passivkonstruktionen hier so gut wie nicht auftreten, wären die Tempora nach den Ergebnissen von PIEPER, die „die Verwendung der einzelnen Tempora in einer Textgruppe in hohem Maße als textgruppenspezifisches Merkmal" ansieht ([30], S. 100), möglicherweise auch sprecherdifferenzierend. Allerdings dürfte die Verwendung von nicht-präsentischen Verbformen bei der verfügbaren Datenmenge pro Sprecher wiederum zu gering für statistisch relevante Aussagen sein: Diese Annahme läßt sich an den Perfektformen belegen, die nach [30], S. 99 in den von PIEPER untersuchten „simuliert gesprochenen Textgruppen" ca. 10% der Verbformen ausmachen, jedoch im VM-Korpus im Mittel nicht einmal eine Häufigkeit von 3 pro Sprecher erreichen, also unter 4% bleiben.

Wie bereits eingangs erwartet, finden sich bei den Sprechern des VM-Korpus zu wenige Imperative für eine statistische Analyse: Während der Imperativ in den „simuliert gesprochenen Textgruppen" PIEPERs der zweithäufigste Modus mit 6% ist, liegen die Werte beim VM-Korpus mit 685 Vorkommen insgesamt, was 0,3% entspricht, im Bereich des Irrelevanten.

Nominale Flexionsformen: Pluralformen, Kasus

MEIER weist auf die Kasusverteilung als interessantes Merkmal hin (vgl. [26], S. 225 ff.), und zwar im Zusammenhang mit seinen sprachstatistischen Untersuchungen an Äußerungen von Zwillingen in Gesprächen, die er dann auf schriftliche Texte von verschiedenen Angehörigen der gleichen Familie ausweitet.

Hier finden sich im Unterschied zu sonstigen Stiluntersuchungen alle Voraussetzungen dieser Arbeit: Persönlicher Stil, spontane Sprache und quantitative Methodik. Doch leider liegen die Kasus im VM-Korpus nicht vor; vielleicht wäre es eine lohnende Arbeit, sie einmal auszuzählen.

Syntax: Teilsätze und Hypotaxen

Die Idee, als weiteres Merkmal die Anzahl der Teilsätze hinzuzuziehen, mußte fallengelassen werden, weil sich die aus den annotierten Bäumen der Dialoge ermittelten Werte teilweise als völlig sinnlos herausstellten[11].

Daß die Unterscheidung zwischen Haupt- und Nebensätzen, die bei PIEPER als Textgruppencharakteristikum eine Rolle spielt (vgl. [30], S. 98), mit den *chunktrees* derzeit nicht möglich ist, fällt dagegen nicht so sehr ins Gewicht, da Hypotaxen in spontansprachlichen Äußerungen ohnehin sehr selten vorkommen: Wenn man von gelegentlichen Relativ- und Kausal- bzw. Konsekutivsätzen absieht, dominieren parataktische Konstruktionen eindeutig.

2.2 Untersuchung der Merkmale

Nachdem in einigen Bereichen klargeworden ist, daß aus Gründen mangelnder statistischer Relevanz der Daten oder schlicht ihrer Nichtverfügbarkeit keine brauchbaren Ergebnisse zu erwarten sind, sollen nun die Untersuchungen an den tatsächlich verfügbaren Merkmalen dargestellt werden.

[11] Beispielsweise ergaben sich bei Zählungen der Nebensatztags häufig weniger Teilsätze, als Sätze aus den Transliterationen der Sprecher! Die *chunktrees* sind einfach noch nicht ausgereift genug, die angegebenen syntaktischen Zerlegungen in den — ironisch gesagt — *„junktrees"* haben keinen allzu großen Realitätsbezug. Hier ist wohl eine neue, verbesserte Version abzuwarten.

2.2.1 Verfügbare Merkmale

Folgende Merkmale aus unterschiedlichen Bereichen der Linguistik wurden in dieser Arbeit näher untersucht[12]:

1. Morphologie: Morphemzahl/Wort
2. Phonologie: Vokale und Konsonanten
3. Lexik
   (a) Wortlängen-Entropien: allgemein und nach Kategorien differenziert[13]
   (b) mittlere Wortlängen: allgemein und nach Kategorien differenziert
   (c) Häufigkeiten kurzer, mittlerer und langer Wörter
   (d) Aussprachevarianten, d.h. umgangssprachliche und dialektale Äußerungen[14]
4. Syntax:
   (a) mittlere Längen von Sätzen und Turns
   (b) Satztypen: Aussage- und Fragesätze
   (c) n-Gramme: Bigramm und Trigramme
   (d) Häufigkeit der Hauptkategorien[15]
5. Semantik: Redundanzen[16]
   (a) „Füllwörter"[17]
   (b) „verbale" und „nonverbale" Verzögerungen[18]

Der Leser mag Merkmale aus den Gebieten der Prosodie und Intonation vermissen, die jedoch eine Untersuchung des Sprachsignals bzw. entsprechene Annotationen erfordern würden; erstere hätte den Rahmen dieser Arbeit gesprengt, letztere hingegen standen nicht zur Verfügung.

[12] Die bereits im vorigen Abschnitt ausgeschlossenen Merkmale tauchen in dieser Aufstellung nicht mehr auf.

[13] Sobald eine Differenzierung nach Kategorien vorgenommen wird, sind die betreffenden Merkmale damit natürlich auch syntaktischer Natur.

[14] Es mag aus linguistischer Sicht bedenklich sein, Umgangssprache und Dialekt zu vermischen, aber zum einen läßt sich beides in der Negation zur Schrift- und Hochsprache zusammenfassen, und zum anderen ist es praktisch unmöglich, diese Unterscheidung aus den VM-Transliterationen zu vollziehen.

[15] sie entstanden durch Zusammenfassung aus den POS-Tags; es handelt sich um ADJ, ADV, AP, ART, KO, N, P, und V sowie OTH für Partikel, Interjektionen u. a.

[16] Eine Aufstellung der strukturalistischen, minimaldistinktiven semantischen Einheit „Sem" oder ähnlicher quantitav erfaßbarer Einheiten ist allerdings in der Praxis nur durch einen vollständigen Parser oder gar Annotationen per Hand erreichbar. Im übrigen sind eindeutige semantische Zuordnungen bei der Ausweitung der ursprünglich ja in der Morphologie entwickelten strukturalistischen Prinzipien auf die Semantik ohnehin nicht möglich. Aus diesen Gründen wurde mit den Redundanzen nur ein sehr begrenzter semantischer Bereich betrachtet, der sich allerdings durch seine spontansprachliche Relevanz auszeichnet.

[17] nach einer relativ willkürlichen Definition des Autors, die auf semantischen Kriterien fußt und die Wörter „also", „dann" und „eigentlich" umfaßt

[18] wiederum zwei relativ willkürliche Definitionen des Autors; erstere sind Häsitationen und Wiederaufnahmen nach Satzabbrüchen, letztere eine Zusammenfassung von akustischen Ereignissen der VM-Transkription (Atmen, Pausen, Zögern und „Schmatzen")

Bevor diese Liste von Merkmalen untersucht werden kann, sollen im folgenden Abschnitt einige Erläuterungen zu den dafür benutzten statistischen Verfahren gegeben werden; der an solchen Ausführungen weniger interessierte Leser wird möglicherweise diesen Abschnitt überspringen und sich direkt der Betrachtung der Merkmale in 2.2.3 zuwenden wollen.

2.2.2 Testverfahren

In der deskriptiven Statistik existieren einige einfache Kenngrößen und Testverfahren, die zur Beurteilung der Frage herangezogen werden können, ob die Verteilung einer Variablen ihre Verwendung in einem Merkmalssatz sinnvoll erscheinen läßt[19].

Mucha nennt als Kriterien zur Merkmalsauswahl und -bewertung Variabilität und Trennstärke (vgl. [28], S. 168). Wir werden zur Bestimmung dieser Kriterien u. a. den von ihm dafür angegebenen Variationskoeffizienten und Histogramme verwenden.

Der Variationskoeffizient ist definiert als Quotient aus Standardabweichung und Mittelwert; er ist ein Maß für die Variabilität einer Stichprobe. Die Einschätzung des Variationskoeffizienten als Variabilitätsmaß einer Verteilung beruht auf der Erwartung, daß sich Mittelwert und Standardabweichung im großen und ganzen proportional verhalten: Je größer der Mittelwert eine Verteilung, desto weiter werden die Werte im allgemeinen um diesen Mittelwert streuen. Ist diese Annahme jedoch in der Weise verletzt, daß die Standardabweichung im Verhältnis zum Mittelwert groß wird, was sich in einem hohen Variationskoeffizienten zeigt, so läßt sich auf eine hohe Variabilität der Verteilung schließen.

Der Grad der stochastischen Abhängigkeit zweier Verteilungen wird durch ihren Korrelationskoeffizienten beschrieben; in dieser Arbeit wird durchgehend der Pearson-Koeffizient als Standardwert eingesetzt. Die Korrelation wird bei der Zusammenstellung der Merkmalssätze verwendet, um Variablen mit zu hoch korrelierten Werten aus dem selben Merkmalssatz auszuschließen[20].

Ein wichtiges grafisches Hilfsmittel der univariaten Statistik sind Histogramme, weil sich mit ihnen eindimensionale Dichten relativ gut approximieren lassen. Insbesondere ein Histogramm mit mehreren ausgeprägten Maxima ist ein gutes Indiz für die Trennfähigkeit des betreffenden Merkmals und eine darin enthaltene Klassenstruktur.

Die bei der Untersuchung der Merkmale verwendeten Tests gehören zu den nichtparametrischen Testverfahren; dies sind Verfahren, die keine bestimmte Verteilung in den Daten voraussetzen, insbesondere keine Normalverteilung. Die Anwendung nichtparametrischer Verfahren begründet sich mit der von Bolz angegebenen Tatsache, daß sprachliche Daten häufig nicht normalverteilt sind (vgl. [7], S. 202). Dabei ist aber die Effizienz nichtparametrischer Verfahren im Vergleich zu ihren parametrischen Gegenstücken hoch genug (96% beim Kolmogorov-Smirnov-Test gegenüber dem t-Test), um eine sinnvolle Verwendung zu gewährleisten; m. a. W., wenn nichtparametrische Verfahren zu keinen signifikanten Ergebnissen führen, werden es parametrische in aller Regel auch nicht tun.

Der χ^2-Test ist wohl der bekannteste und einfachste Anpassungstest einer theoretischen Verteilung an empirisch ermittelte Werte; er ist für diskrete Verteilungen absoluter Gesamtmengen mit möglichst 5 oder mehr Beobachtungen pro Klasse geeignet (vgl. [29], S. 119).

---

[19] Die in dieser Arbeit verwendeten Tests und Kenngrößen sowie die dazugehörigen Grafiken wurden mit den Programmen *MS Excel* und *Systat* für Windows auf einem handelsüblichen PC erstellt.

[20] s. aber die Diskussion in 2.3.1

MULLER verwendet ihn u. a. zum Testen der Verteilungen grammatischer Kategorien und des Vokabulars.

Er beruht auf der Teststatistik

$$\chi^2 = \sum \frac{(w - th)^2}{th} \quad \text{mit den Freiheitsgraden} \quad \nu = (n-1)(k-1) \quad .$$

w und th sind dabei die empirischen bwz. theoretischen Werte, n und k die Zeilen bzw. Spalten der Tabelle, die aus diesen Werten gebildet wird.

Bei hinreichender Größe führt die Teststatistik zur Ablehnung der Nullhypothese, daß die empirische Verteilung mit der sich aus den theoretischen Gesamtmengen ergebenden Verteilung übereinstimmt; die Ablehnung erfolgt auf einem vorher festzusetzenden Signifikanzniveau (für das übrigens für diese Arbeit stets der allgemein übliche Wert 5% gewählt wurde).

Dieser Test wird bei den folgenden Untersuchungen auf zweierlei Weisen angewendet, und zwar „vertikal" und „horizontal" in Bezug auf die jeweiligen Datenmatrizen[21]. Im ersten Fall fungiert die Gesamtmenge an vorhandenem Text aller Sprecher als Grundgesamtheit, die durch die einzelnen Sprecher unterteilt wird. Die theoretischen Gesamtmengen ergeben sich als Bruchteile des untersuchten Merkmals zu einer Referenzgröße, die die Textlänge charakterisiert (etwa die Wortzahl für Kategorien); die Nullhypothese ist die gleichmäßige Verteilung eines Merkmals über die Gesamtheit der Sprecher. Kann man die Hypothese auf einem geeigneten Signifikanzniveau ablehnen, so erweist sich das betrachtete Merkmal als signifikant abweichend von der durch die Gesamtmege vorgegebenen Verteilung.

Im anderen Fall wird die Verteilung eines Merkmals bei einem einzigen Sprecher untersucht[22]. Man bildet dazu eine zufällige Stichprobe gleicher Größe aus dem gesamten Textmaterial aller Sprecher und betrachtet die Verteilung innerhalb dieser Stichprobe als theoretische Werte für den Test. Kann man hier die Nullhypothese ablehnen, so erweist sich der Sprecher bezüglich des Merkmals als signifikant abweichend von der Gesamtheit der Sprecher; findet man „genügend" solcher Abweichungen, so erweist sich das Merkmal wiederum als ausreichend differenziert.

Neben dem χ^2-Test kommt als weiteres nichtparametrisches Verfahren der KOLMOGOROV-SMIRNOV-Test zur Anwendung. Seine Teststatistik $K_{m,n}$ ist im 1-Stichproben-Fall die maximale Differenz zwischen einer empirischen und einer theoretischen Verteilungsfunktion und im 2-Stichproben-Fall die Differenz zweier empirischer Verteilungsfunktionen $F_m(z)$ und $G_n(z)$ mit m bzw. n Beobachtungen (vgl. [9], S. 133 ff.):

$$K_{m,n} = \max_z |F_m(z) - G_n(z)|$$

Erreicht die Teststatistik einen ausreichend hohen Wert, so wird die Nullhypothese der Gleichverteilung beider betrachteteter Verteilungen abgelehnt. Der Test ist bereits für eine geringe Zahl von Beobachtungen, wie sie bei vielen Merkmalen des VM-Korpus vorliegen, gut geeignet; allerdings geht er von einer stetigen Veteilungsfunktion der Variablen aus, was bei den untersuchten Merkmalen wiederum oft nicht der Fall ist. In diesen Fällen wird er „konservativ", verliert also etwas an Aussagekraft.

---

[21] Die Datenmatrix (der Begriff wird bei der Clusteranalyse wieder begegnen) enthält als Zeilen die Einzelbeobachtungen, also die Sprecher, und als Spalten die verschiedenen Merkmale.

[22] Dieser Fall entspricht im wesentlichen der bei [29] angegebenen Verwendung des χ^2–Tests für Textstichproben.

In der 1-Stichproben-Version erfüllte der KOLMOGOROV-SMIRNOV-Test lediglich die „routinemäßig“ unterstützende Aufgabe, die o. a. Annahme in [7], S. 202 zu bestätigen, daß bei sprachlichen Merkmalen häufig nicht mit Normalverteilung zu rechnen ist, was ja gerade die Anwendung nichtparametrischer Verfahren erforderlich macht. Dazu wurden die Merkmale mit der z-Transformation, die auch später bei der Klassifikation zur Anwendung kommt,

$$z = \frac{x - \mu}{\sigma} \quad \text{(mit dem Mittelwert } \mu \text{ und der Standardabweichung } \sigma \text{)}$$

auf Mittelwert 0 und Standardabweichung 1 standardisiert und anschließend gegen die Standardnormalverteilung getestet. Im Ergebnis konnte die Nullhypothese „das (standardisierte) Merkmal XY ist (standard)normalverteilt“ in der Tat fast generell mit 5% Signifikanz abgelehnt werden, so daß darauf bei den einzelnen Merkmalen nicht weiter eingegangen wird: Die nichtparametrischen Verfahren sind für diese Arbeit die Methode der Wahl und bleiben es auch dann, wenn das eine oder andere Merkmal normalverteilt sein sollte.

Von größerer Bedeutung für die vorliegende Arbeit ist die 2-Stichproben-Version des KOLMOGOROV-SMIRNOV-Tests.

Sie wird zur Erkennung signifikanter Abweichungen bei den Merkmalen verwendet, die keinen χ^2-Test zulassen, weil sie relative Häufigkeiten oder selbst theoretische Gesamtmengen für χ^2-Tests darstellen. Dabei wird für einen Sprecher die empirische Verteilung aus den Häufigkeiten des fraglichen Merkmals, bezogen auf die jeweilige Basiseinheit (Satz oder Turn), gebildet. Diese Verteilung wird gegen die Verteilung der jeweiligen Variablen bei einem anderen Sprecher getestet. Stellt sich dabei in hinreichend vielen Fällen eine signifikante Differenz ein, so kann die getestete Variable als trennstarkes Merkmal gelten.

Bei der Zusammenstellung der Merkmalsätze schließlich wird mit dem 2-Stichproben-KOLMOGOROV-SMIRNOV-Test die Verteilung je zweier in Frage kommender Merkmale gegeneinander getestet; wenn ihre Verteilungen sich nicht signifikant voneinander unterscheiden, werden sie als ungeeignet für die Verwendung im selben Merkmalssatz angesehen. Auch hier wurden die Merkmale vor dem Test standardisiert.

2.2.3 Geeignete Merkmale

Bevor die in 2.2.1 angeführten Merkmale mit den vorgestellten Verfahren im einzelnen auf ihre Tauglichkeit in Merkmalssätzen untersucht werden, stellen wir zunächst noch einige Betrachtungen an, die begründen sollen, warum die Einführung von Basisgrößen erforderlich ist.

Turnwerte und die „Norm“

Natürlich ist es sehr aufwendig, alle Sprecher der Trainingsmenge auf die im vorhergehenden Abschnitt angegebene Weise gegeneinander zu testen[23]. Dies mag für die Differenzierung von einigen Textgruppen oder die Unterscheidung zweier Autoren ein praktikables Verfahren sein; für die Trainingsmenge wird in dieser Arbeit statt dessen eine „Norm“ erstellt.

Dazu wurde eine Zufallsstichprobe mit der gleichen Turnzahl wie die Daten des Sprechers mit der maximalen Turnzahl aus den Daten aller Sprecher gebildet. Gegen diese Stichprobe als Referenz werden dann die einzelnen Sprecher, genauer gesagt die Verteilungen des jeweils untersuchten Merkmals bei den einzelnen Sprechern, getestet. Damit die tatsächlichen und

[23] bei 470 Sprechern wäre jedesmal eine Matrix mit über 2 Millionen Einträgen zu betrachten

erwarteten Werte bei den horizontalen χ^2-Tests von der Größenordnung her vergleichbar sind, wurden die Normwerte auf die Gesamtsumme des jeweils untersuchten Sprechers skaliert, so daß die Summe der Normwerte jedesmal gleich der Summe der Sprecherwerte ist[24].

Generell sind bei stark unterschiedlichen Textlängen, wie sie im VM-Korpus bei den einzelnen Sprechern vorliegen, keine absoluten Häufigkeiten verwendbar: Sie werden stets eine hohe Korrelation zur Gesamt-Textlänge aufweisen und sollten daher besser auf eine Einheitsgröße bezogen werden. Hierfür wurden in dieser Arbeit drei Varianten betrachtet.

Zum ersten wurde wie in [30] der Satz, die übliche Basisgröße für quantitative Stilmerkmale, verwendet, weil ein Verzicht auf so charakteristische Variablen wie die Satzlänge nicht statthaft erschien und sich bei ausreichend häufiger Ausprägung eines Merkmals Vergleiche von Satz- und Turnwerten anbieten. Jedoch muß man sich darüber im klaren sein, daß Sätze primär als Untersuchungseinheit für geschriebene Sprache ihre Berechtigung haben[25].

Für gesprochene dialogische Sprache hingegen ist eher der Turn als sinnvolle Basiseinheit anzusehen[26]. Die Vorteile gegenüber dem Satz als Grundeinheit liegen in der besseren statistischen Verwendbarkeit durch höhere Einzelwerte, weil ein Turn oftmals mehrere Sätze umfaßt, sowie in der höheren linguistischen Motivation[27].

Schließlich lassen sich, um der unterschiedlichen Textlänge Rechnung zu tragen, relative Häufigkeiten verwenden. Diese haben den Vorteil, nach Betrachtung einer beliebig kurzen Textmenge zur Verfügung zu stehen, wenn auch ihre Aussagekraft erst mit einem längeren Text gegeben ist.

Aktionsquotienten

Ein typisches stilistisches Merkmal geschriebener Sprache ist der sog. Aktionsquotient[28]. Mit diesem Quotienten aus der Anzahl der finiten Verben und der attributiven Adjektive unter Einschluß partizipialer Formen läßt sich mit einer einleuchtenden, wenngleich nicht ganz unproblematischen linguistisch-psychologischen Interpretation eine Untergliederung von Texten in „aktive" und „deskriptive" vornehmen. Jedoch ergibt sich beim Versuch, diese Einteilung auf die im VM-Korpus enthaltenen Dialoge anzuwenden, nicht ganz unerwartet der Befund, daß sämtliche Textproben der einzelnen Sprecher als „aktiv" einzustufen sind: Dialoge geben „deskriptiven" Elementen eben keinen Raum.

Der Aktionsquotient ist somit eigentlich schon durch diese einfachen Überlegungen für gesprochene Sprache diskreditiert. Das Resultat wird aber auch durch den von ALTMANN angegebenen FISCHER-Test ([2], S. 29 ff.) für den Vergleich zweier Aktionsquotienten mit der z-Statistik

[24] Dieses Verfahren entspricht im Prinzip der in [30], S. 95 geschilderten Umrechnung der absoluten in relative Häufigkeiten.

[25] PIEPER führt für schriftliche Texte stichhaltige theoretische und praktische Gründe an, vgl. [30], S. 34

[26] Ein Turn kann im Extremfall ausschließlich aus nonverbalen Produktionen bestehen; rein technisch enthält er dann aber für die Zählungen dieser Arbeit trotzdem einen Satz, sofern in der Transliteration eine entsprechende Interpunktion gesetzt wurde.

[27] Bei der Transliteration spontansprachlicher Äußerungen zeigt sich die Problematik des Satzbegriffs: Man muß sich bei der Interpunktion für eine Lesart entscheiden, obwohl häufig Alternativen denkbar wären — man denke nur an Satzabbrüche, Wiederholungen, Parenthesen u. a. Turngenzen hingegen lassen sich deutlich leichter und fast stes eindeutig festmachen (wenn man einmal von Unterbrechungen durch den Gesprächspartner absieht).

[28] vgl. [2], S. 18 ff.; ich verwende allerdings aus technischen Gründen die Version von [30], S. 68 ff.

$$z = \frac{Q_1 - Q_2}{\sqrt{Q_1 Q_2 (1/v_1 + 1/a_1 + 1/v_2 + 1/a_2)}}$$

bestätigt; Q_1 und Q_2 sind dabei die beiden fraglichen Aktionsquotienten, v_1, v_2 die Verb- und a_1, a_2 die Adjektivhäufigkeiten der beiden untersuchten Texte.

Diese z-Statistiken wurden im paarweisen Vergleich jedes Sprechers mit der oben eingeführten Norm errechnet; sie fungieren als Quantile einer zweiseitigen Normalverteilung und erreichen dabei lediglich für 32 von 470 Sprechern, also in 6,8% der möglichen Fälle den kritischen Wert für 5% Signifikanz. Der Aktionsquotient kommt damit als Merkmal nicht in Frage.

Sonstige Quotienten: Clauselänge, Dichtegrad

In [30], S. 62 ff. werden neben dem Aktionsquotienten noch weitere Quotienten betrachtet, nämlich Clauselänge und Dichtegrad. Jedoch ergibt ihre Untersuchung beim VM-Korpus keine hohe Trennfähigkeit, wie wir gleich sehen werden.

Clauselängen sind definiert als die Anzahl der Wörter, die sich um ein finites Verb gruppieren. Ihre Mittelwerte bei den einzelnen Sprechern liegen im Bereich von 8,65 mit einer Standardabweichungvon 1,0, sie haben also mit einem Variationskoeffizienten von 0,12 einen geringe Variabilität.

Nimmt man das Ergebnis des KOLMOGOROV-SMIRNOV-Tests hinzu, der — auf Turnbasis durchgeführt — für 55 von 470 Sprecher Abweichungen von der Norm mit 5% Signifikanz zeigte, d.h. in lediglich 11,7% der Fälle, so läßt sich im Ergebnis feststellen, daß Clauselängen beim VM-Korpus kein trennstarkes Merkmal darstellen.

Wie verhält es sich nun mit dem Dichtegrad?

Dieses Merkmal ist ein relativ komplex zusammengesetzter Quotient; ihm liegt die psycholinguistische Auffassung zugrunde, daß die Aussage eines Satzes primär von attributiven Adjektiven, Nomina und finiten Verben getragen wird. Er ist definert als Quotient aus der Summe dieser Kategorien und der Wortzahl. Im VM-Korpus liegen die mittleren Dichtegrade der einzelnen Sprecher im Durchschnitt bei 0,26, bei einem recht niedrigen Variationskoeffizienten von 0,13.

Beim KOLMOGOROV-SMIRNOV-Test ergab sich in 70 von 470 Fällen, also 14,9%, eine signifikante Abweichung von der Norm, mithin ein ähnliches Ergebnis wie bei den Clauselängen.

Im Histogramm (Abb. 2.1, S. 17) zeigen sich zwar kleinere Abweichungen innerhalb des einen Maximums, aber von mehreren ausgeprägten Maxima kann nicht die Rede sein.

Zusammen mit den Ergebnissen des vorigen Abschnitts können wir also festhalten, daß keiner der von PIEPER betrachteten, aus Kategorien und Wortzahlen zusammengesetzten Quotienten als Merkmal für die stilistische Klassifikation der spontansprachlichen VM-Daten in Frage kommt.

Kategorien

Um ausreichende Häufigkeiten auch bei Sprechern mit wenig Text zu erhalten, wurden die Kategorien des STTS-Taggers zu neun Hauptkategorien zusammengefaßt. Die so erhaltene Häufigkeitsverteilung wurde zunächst mit dem oben beschriebenen Verfahren des horizontalen χ^2-Tests gegen die Norm getestet. Es ergab sich in 212 von 470 Fällen Signifikanz auf dem

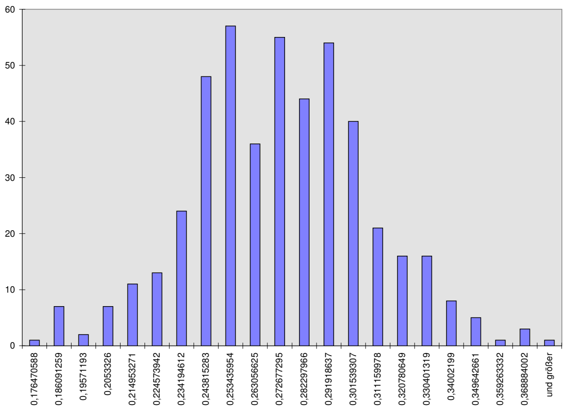

Abbildung 2.1: Histogramm der mittleren Dichtegrade

Niveau 5%, was immerhin 45% der Fälle entspricht. Die Verteilung der Hauptkategorien scheint also bei den einzelnen Sprechern recht unterschiedlich zu sein.

Um nun den Beitrag der einzelnen Kategorien zu dieser Verschiedenheit zu überprüfen, wurden mit vertikalen χ^2–Tests die Verteilungen der einzelnen Kategorien über die Gesamtheit der Sprecher überprüft. Dabei zeigten sich signifikante Abweichungen in der Verteilung gegenüber den theoretischen Gesamtsummen (also der Wortzahl) bei fast allen Kategorien, insbesondere bei ADV und AP. Verben bilden eine Ausnahme: Sie verteilen sich nach dem χ^2-Resultat (der Wert für die Wahrscheinlichkeit liegt sehr nahe bei 1) fast genau so auf die Sprecher wie die Wortzahlen insgesamt; die relative Häufigkeit von Verben kommt damit als Merkmal nicht in Frage.

Von den Hauptkategorien treten N, P und V häufiger auf als ADV und AP, sind aber weniger stark gestreut, ART und KO hingegen weisen einen höhereren Variationskoeffizienten auf, haben aber mit 5% und 3% zu geringe relative Häufigkeiten.

Das Histogramm der relativen ADV-Häufigkeiten (s. Abb. 2.2, S. 18) weist nur ein Maximum auf; ähnlich verhält es sich mit den Histogrammen der anderen Kategorien, die wir aus diesem Grund nicht aufführen. Dennoch wird die relative ADV- und AP-Häufigkeit mangels besserer Alternativen und aufgrund des χ^2-Testergebnisses als kategoriales Merkmal im Auge behalten.

Ein anderes Bild hinsichtlich der Differenzierfähigkeit der Kategorien ergibt sich, wenn wir die Turns mit einbeziehen: Hier wurden N, P und V wegen des ähnlichen Variationskoeffizienten ihrer turnbezogenen Mittelwerte (ca. 0,4) und ihrer höheren Häufigkeit pro Turn untersucht[29]. Die Turnwerte der einzelnen Sprecher wurden gesammelt und die dadurch entstehenden empirischen Verteilungen gegen die Verteilung der entsprechenden Werte in der Norm getestet. Dabei ergab sich für die Nomina in 29,6% der Fälle eine 5%-signifikante Abweichung von der Norm. Für die Kategorie P betrug die Anzahl der Signifikanzen 37,2%, während für die Verben mit 191 von 470 Fällen, was 40,6% entspricht, der beste Wert erreicht wurde. Als Ergebnis bleibt festzuhalten, daß bei den relativen Häufigkeiten ADV und

[29] Die Häufigkeiten der anderen Kategorien pro Turn liegen deutlich niedriger als die Werte dieser Kategorien; sie wurden deshalb nicht weiter betrachtet.

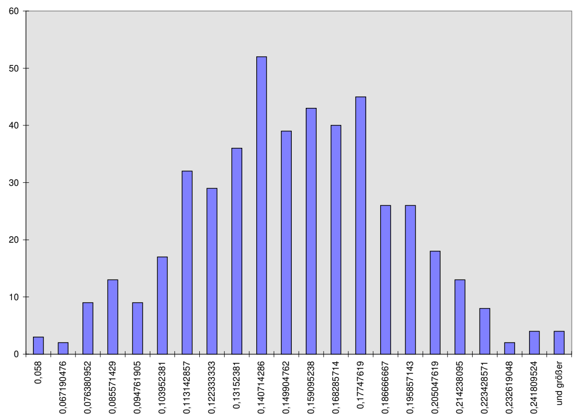

Abbildung 2.2: Histogramm der relativen ADV-Häufigkeiten

AP und bei den turnbezogenen Werten V die brauchbarsten Variablen sind und diese in Merkmalssätzen verwendet werden können.

Schließlich wurden die relativen Häufigkeiten von Kategorien am Satzende betrachtet — mit dem linguistisch motivierten Hintergedanken, daß das Satzende oftmals eine besonders herausgehobene Position im Satz innehat. Hierbei ergaben sich für V und ADJ relative Häufigkeiten von 0,2 bei einem recht guten Variationskoeffizienten 0,5; diese beiden Merkmale werden uns weiter unten noch beschäftigen.

Wortlängenmerkmale

Die Wortlängen wurden allgemein und nach den eben vorgestellten Haupkategorien differenziert ausgewertet.

Zunächst wurden die mittleren Wortlängen pro Satz bestimmt und die so entstehenden empirischen Verteilungen der einzelnen Sprecher wieder mit dem KOLMOGOROV-SMIRNOV-Test gegen die Norm getestet. Dabei ergaben sich durchgängig signifikante Abweichungen; die Verteilung der Wortlängen bei den einzelnen Sprechern ist also höchst heterogen.

Wenn wir jedoch die sprecherbezogenen Mittelwerte betrachten, sind die Resultate weniger günstig: Die größte Variabilität ergibt sich dabei für die Adjektive, die einen Variationskoeffizienten von knapp 0,14 erreichen, während die Wortlängen der anderen Kategorien und die allgemeinen Wortlängen unter 0,1 liegen.

Falls man aber auf kategoriebezogene Wortlängenmerkmale nicht verzichten will, stellen also die ADJ-Wortlängen dafür noch die beste Möglichkeit dar.

Nachdem der Mittelwert der Wortlängen bei 5 liegt, wurde neben der Differenzierung über die Hauptkategorien noch eine andere versucht, und zwar die Unterteilung in Wörter kurzer (< 4), mittlerer ($4 - 5$) und langer Wortlänge (> 5). Dabei werden also nicht mehr Wortlängen und Mittelwerte daraus als Variablen betrachtet, sondern die Verteilung der unterschiedlichen Wortlängengruppierungen bei den Sprechern. Die zugrundeliegende Idee ist, daß manche Sprecher möglicherweise häufiger lange Wörter verwenden als andere.

Zunächst wurden die Variablen horizontalen χ^2-Tests unterzogen. Nur in 78 von 470 Fällen, d. i. 16,6%, waren die Abweichungen signifikant.

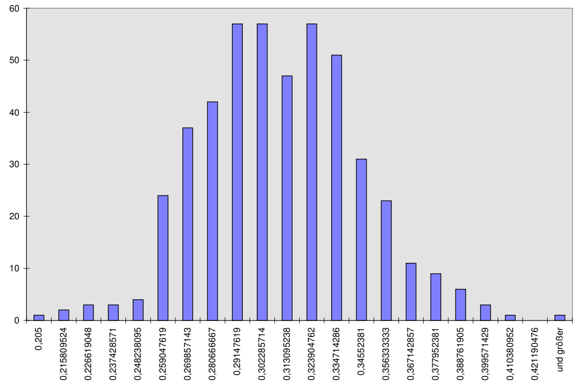

Abbildung 2.3: Histogramm der relativen Häufigkeiten langer Wörter

Um zu überprüfen, ob dieses Ergebnis für alle drei Wortlängengruppen gilt, wurden auch diesmal vertikale χ^2-Tests durchgeführt. Dabei stellte sich heraus, daß die Wörter mit kurzen Wortlängen laut χ^2-Wert genau so auf die Sprecher verteilt sind wie die Wortzahlen insgesamt — also ein ähnliches Ergebnis wie oben bei den Verben. Bei der Anzahl von Wörtern mittlerer und langer Wortlängen hingegen zeigten sich signifikante Abweichungen von der durch die totale Wortzahl vorgegebenen theoretischen Gesamtmenge. Nun sind mittlere und lange Wörter durchaus seltener als kurze, das Verhältnis ist aber mit 40:30:30 nicht so eindeutig, daß man die höchst unterschiedlichen χ^2-Resultate allein damit erklären könnte. Die Signifikanzwerte sind jedoch bei beiden Gruppen deutlich geringer als etwa bei den oben betrachteten Haupkategorien ADV und AP.

Hinzu kommt, daß die Variationskoeffizienten der relativen Häufigkeiten von Wörtern mittlerer und langer Wortlänge nur etwa 0,1 betragen, bei den kurzen Wörtern sogar noch darunter; immerhin ein weiteres Indix, daß die geringen Schwankungen in den Verteilungen, von mittleren und langen Wörtern verursacht werden, wie sich ja auch bei den χ^2-Tests gezeigt hat. Wir führen exemplarisch das Histogramm der relativen Häufigkeiten langer Wörter an (Abb. 2.3), das bis auf die leichte Irregularität in der Mitte der Verteilung lediglich ein einzelnes deutliches Maximum zeigt.

Zusammenfassend läßt sich also feststellen, daß die Häufigkeiten von Wörtern mittlerer und langer Wortlänge, seien es nun relative oder turnbezogene, mit Einschränkungen als Merkmale geeignet, die kurzen Wörter aber eindeutig abzulehnen sind.

Wortlängen-Entropien

ALTMANN benutzt in [2], S. 38 ff. die aus der Informationstheorie bekannte Entropie

$$H_1 = -\sum_i p_i \log p_i \quad \text{mit} \quad p_i = \frac{f_i}{N} \quad \text{und} \quad N = \sum_i f_i$$

(wobei die Wahrscheinlichkeit p_i der i-ten Texteinheit durch die relative Häufigkeit f_i/N abgeschätzt werden kann) als Maß für die globale Diversität von Textuntereinheiten im Gesamttext; er bezieht die Entropie dabei auf Wortlängen und gibt ein Kriterium für den t-Test zum Vergleich zweier Texte an ([2], S. 43):

$$t = \frac{H_1 - H_2}{\sqrt{varH_1 + varH_2}} \quad \text{mit den Freiheitsgraden} \quad FG = \frac{(varH_1 + varH_2)^2}{\frac{(varH_1)^2}{N_1} + \frac{(varH_2)^2}{N_2}}$$

Dieses Kriterium wurde aus der paarweisen Zusammenstellung der Wortlängen-Entropien der einzelnen Sprecher und der Norm ermittelt und diente dann als Quantil einer t-Verteilung mit den berechneten Freiheitsgraden.

Für die allgemeinen Wortlängen zeigen sich zunächst keinerlei signifikante Werte. Das Bild ändert sich jedoch grundlegend, wenn man nach Kategorien differenziert: Hier ergibt sich für die ADJ-Wortlängen-Entropie das beste Ergebnis, und zwar in 389 von 470 Fällen eine 5%ige Signifikanz. Als anschauliche Erklärung dafür läßt sich anführen, daß auch der Variationskoeffizient bei adjektivischen Wortlängen am höchsten ist: Diese weisen die höchste Variabilität unter den kategoriebezogenen Wortlängen auf, und so ist es nur plausibel, daß auch ihre Diversität am größten ist.

Auch wenn die Variabilität der ADJ-Wortlängen-Entropie recht niedrig ist (der Variationskoeffizient liegt nur bei 0,11), halten wir sie aufgrund der Testresultate als trennstarkes Merkmal für die weiteren Untersuchungen fest[30].

Sonstige Intra-Wort-Merkmale: Vokale/Konsonanten, Morpheme, Silben

Die Verteilung von Vokalen und Konsonanten, ohnehin eine Merkmalsgruppe, die eher für poetische Texte Anwendung findet[31], erweist sich durch den „vertikalen“ χ^2–Test sehr schnell als ungeeignet für gesprochene Sprache: Es ergeben sich keinerlei signifikante Abweichungen von der Buchstabensumme im Text als theoretischer Gesamtwert. Demzufolge zeigt auch die Variable „relativer Anteil von Vokabeln“ eine äußerst geringe Standardabweichung und einen sehr niedrigen Variationskoeffizienten von knapp 0,025. Der Vokalanteil liegt fast konstant bei einem Wert von 0,36 und unterschreitet damit leicht das von Meier in [26], S. 248 angegebene Vokal/Konsonanten-Verhältnis von 40/60. — Man kann also mit ziemlicher Sicherheit behaupten, daß die Sprecher im VM-Korpus sich weder durch eine besonders vokalreiche noch eine besonders vokalarme Sprache auszeichnen.

Bei den Morphemen ist zunächst festzustellen, daß die mittlere Morphemlänge wohl zu Doležels L-Charakteristiken (vgl. [12], S. 18) zu rechnen ist, also nur von der Sprache allgemein, nicht aber vom jeweiligen Stil abhängt. Sie ist damit ungeeignet zur Sprecherdifferenzierung, was auch durch dieStatistik bestätigt wird: Die Morphemlänge liegt im Mittel bei etwas über 3 mit einer Standardabweichung von nur 0,06.

Es kann nicht weiter überraschen, daß die Anzahl der Morpheme pro Wort stark mit der Wortlänge korreliert (der Wert liegt bei 0,75); weil aber diese leichter zu ermitteln ist, kann auf die Berücksichtigung der mittleren Morphemzahl verzichtet werden.

[30]

[31] wie auch [2] sich bei Vokalwiederholungen v. a. mit poetischen Texten befaßt

Als weiteres Intra-Wort-Merkmal könnte die (Sprech-)Silbenzahl von Interesse sein, wenngleich GROTJAHN für dieses Merkmal v. a. eine textgruppendifferenzierende (zwischen Poesie und Prosa) sowie eine situationsdifferenzierende Funktion (zwischen privater und geschäftlicher Korrespondenz) ermittelt (vgl. [17], S. 173 ff.). Sie läßt sich aus den Transliterationen des VM-Korpus jedoch nicht direkt ermitteln und steht ohnehin in sehr engem Zusammenhang mit der Wortlänge, weshalb in dieser Arbeit die einfach zu ermittelnden Wortlängenmerkmale bevorzugt wurden.

Satztypen

An Satztypen finden sich im VM-Korpus, sei es als Folge der Transliteration oder aufgrund der Gesprächssituation, keinerlei Ausrufe, sondern lediglich Interrogativsätze und Aussagen.

Für Aussagesätze ergab sich beim vertikalen χ^2–Test keine Signifikanz, für Fragesätze jedoch war sie sehr hoch. Dies bedeutet, daß Fragesätze von der durch die Gesamtzahl der Sätze gegebenen Verteilung der theoretischen Gesamtwerte signifikant abweichen. Auch ihr Variationskoeffizient liegt mit 0,51 deutlich über dem der der Aussagesätze, der nur 0,1 beträgt. Demzufolge wäre die Variable „relative Häufigkeit von Fragesätzen“ als Merkmale in die weiteren Untersuchungen mit einzubeziehen.

Man kann aber nicht davon ausgehen, daß von zwei Variablen, die funktional voneinander abhängen, die eine trennstark ist und die andere nicht. Und so läßt sich das Resultat aus dem χ^2–Test denn auch damit begründen, daß der relative Anteil von Fragesätzen im Vergleich zu Aussagen gering ist und sich kleinere „Ausreißer“ damit beim Test stärker bemerkbar machen. Der höhere Variationskoeffizient der Fragesätze gegenüber den Aussagen erklärt sich aus der geringeren relativen Häufigkeit. die im Schnitt bei 0,17 liegt, wobei die Standardabweichung natürlich bei beiden Satztypen mit 0,8 den gleichen Wert annimmt.

Wir müssen demnach als Ergebnis festhalten, daß Satztypen keine trenn- und variabilitätsstarken Variablen darstellen und damit nicht für Merkmalssätze in Frage komen.

Satz- und Turnlängen

Die mittleren Satz- und Turnlängen (letztere in Sätzen gemessen) sind fundamentale Merkmale geschriebener bzw. gesprochener Sprache. Im VM-Korpus weisen diese beiden Variablen einen nicht sehr hohen, aber auch nicht unbrauchbar niedrigen Variationskoeffizientvon knapp über 0,3 auf.

Beim 2-Stichproben-KOLMOGOROV-SMIRNOV-Test ergaben sich für die Satzlängen bei 156 von 470 Sprechern 5%-signifikante Werte, also immerhin in einem Drittel der Fälle. Noch deutlicher waren die Resultate bei den Turnlängen: Hier waren die Werte für 265 von 470 Sprechern auf dem 5%-Niveau signifikant verschieden von den durch die Norm vorgegebenen Werten, was einer Quote von 56% entspricht.

Die Histogramme (s. Abb. 2.5, S. 27 im Abschnitt 2.3.2) haben eine gewisse Ähnlichkeit in ihrer linkssteilen Form, die Merkmale sind aber nicht korreliert. Ihre standardisierten Verteilungen zeigen beide im 1-Stichproben-KOLMOGOROV-SMIRNOV-Test eine derart signifikante Abweichung von der Standard-Normalverteilung, daß sie als nicht normalverteilt gelten können. Dies gilt im übrigen auch für ihre logarithmierten Werte, so daß die Ergebnisse von BUCH in [8] bestätigt werden, der ebenfalls den nichtparamatrischen KOLMOGOROV-SMIRNOV-Test auf Satzlängenverteilungen anwendet, weil die Hypothese der Normalverteilung abzulehnen ist.

Aufgrund ihrer Variabilität, Trennfähigkeit und ihrer grundsätzlichen linguististischen Relevanz und Bedeutung für die Stilanalyse werden diese Merkmale in Merkmalssätze aufgenommen.

Spontansprachliches: Aussprachevarianten, "Verzögerungen", „Füllwörter"

Kommen wir nun zum fast ausschließlich spontansprachlich relevanten Bereich. Zu Beginn diese Kapitels wurde bereits darauf hingewiesen, daß im VM-Korpus neben dem reinen Text auch Häsitationen, Aussprachevarianten und nonverbale Produktionen transliteriert sind.

Es ist nun schwierig, für nonverbale Produktionen eine theoretische Gesamtmenge für den χ^2-Test festzulegen, weil sie nicht in die Wortzahl eingehen, die daher bestenfalls ein Schätzwert für die theoretische Gesamtmenge sein kann. Auf χ^2-Tests wurde daher in diesem Abschnitt verzichtet.

Aus linguistischer Sicht lassen sich Häsitationen und Wiederaufnahmen nach Satzabbrüchen zu verbalisierten redundanten Verzögerungen im Ablauf der Textproduktion zusammenfassen. Dies wurde im Rahmen der vorliegenden Arbeit auch unter dem Gesichtspunkt der geringen Häufigkeit von Wiederaufnahmen[32] getan und das entstehende Merkmal wird im folgenden mit „verbale Verzögerungen" bezeichnet.

Ähnlich wurde mit den nonverbalen Produktionen Atmen, Pausen, Zögern und „Schmatzen" verfahren, die ebenfalls Redundanzen darstellen und den Dialog verzögern: Sie wurden zu den „nonverbalen Verzögerungen" zusammengefaßt[33], was vor allem im Hinblick auf den Satz, aber auch auf den Turn als Basiseinheit geboten war.

In den Bereich der Lexik sowie der Semantik gehören die „Füllwörter"; sie werden aber trotzdem in diesem Abschnitt aufgeführt, weil ihre Häufung gerade für spontane Sprache charakteristisch ist. Mit einem Mittelwert von unter 1 pro Turn und einem Variationskoeffizienten von 0,76 liegen sie im unteren Bereich der in diesem Abschnitt betrachteten Merkmale; der geringe Mittelwert läßt die Anwendung eines KOLMOGOROV-SMIRNOV-Tests nicht zu.

Die Aussprachevarianten ließen sich, obgleich seltener anzutreffen als etwa Zögern, nicht in dieses Schema der Redundanzen integrieren und werden deshalb gesondert betrachtet. Sie weisen bezogen auf Sätze und Turns einen gleichermaßen hohen Variationskoeffizienten von etwa 1 auf, d.h. ihre Standardabweichung ist ebenso hoch wie ihr Mittelwert, der für Turns ebenfalls bei etwa 1 liegt. Der hohe Variationskoeffizient ist vor allem auf die große Anzahl von Sätzen und auch Turns zurückzuführen, in denen keinerlei Aussprachevarianten vorkommen, wie auch das zusammenfassende Histogramm der satz- und turnbezogenen sponntansprachlichen Werte zeigt (Abb. 2.4, S. 23).

Anders die Häsitationen: Ihr Histogramm weist eine große Anzahl von relativen Häufigkeiten nahe bei 0 aus, aber nur wenige tatsächliche Nullwerte, während der Variationskoeffizient ebenso hoch ist wie bei den Aussprachevarianten.

Betrachten wir nun die Verzögerungen: Die verbalen Verzögerungen verhalten sich hinsichtlich Mittelwert (bezogen auf Turns liegt er bei etwa 0,9 und bezogen auf Sätze bei 0,4) und Variationskoeffizient ähnlich wie die Aussprachevarianten, zeigen aber im Histogramm starke Analogien zu den Häsitationen (was aufgrund ihrer Definition nicht verwundert); sie stellen also in ihren Eigenschaften hinsichtlich Variabilität und Trennstärke eine Mischung zwischen diesen beiden Merkmalen dar. Bei den nonverbalen Verzögerungen ergibt sich, daß

[32] durchschnittlich knapp 5 Vorkommen pro Sprecher

[33] Unter den darin enthaltenen Merkmalen sind Pausen und Schmatzen mit Durchschnittswerten von 7 bis 8 pro Sprecher die wirklich seltenen, während Atmen und Zögern häufiger zu finden sind.

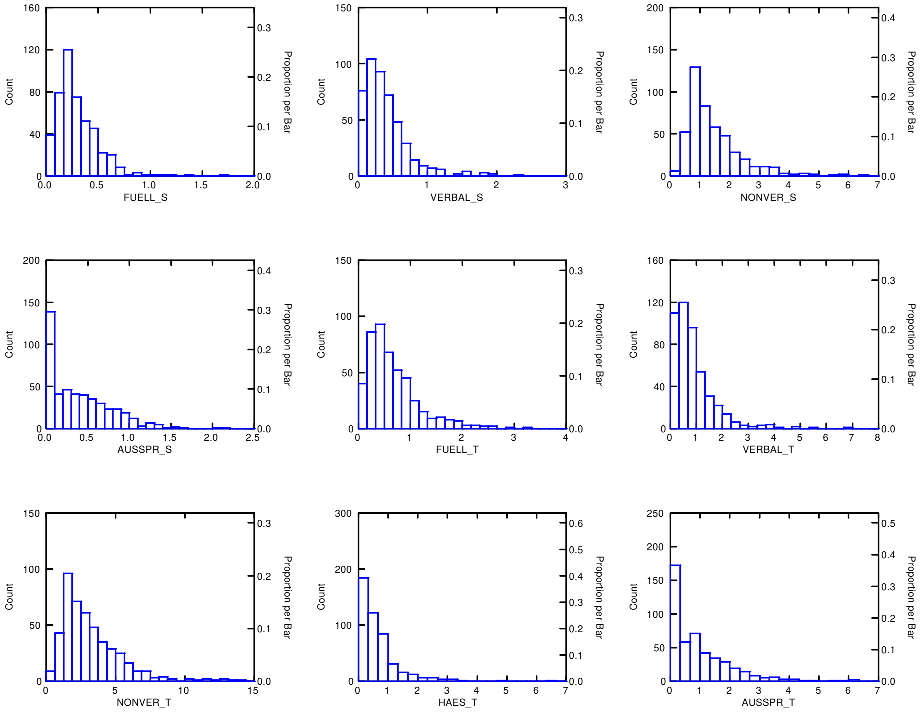

Abbildung 2.4: Histogramme spontansprachlicher Merkmale

sie den niedrigsten Variationskoeffizienten der in diesem Abschnitt untersuchten Merkmale aufweisen, der sich aber aus ihrer höheren Anzahl erklärt (sie liegt über 3 pro Turn).

Wir können also zusammenfassend feststellen, daß sich alle betrachteten spontansprachlichen Variablen bzw. die daraus definierten Summen durch hohe Variabilität auszeichnen, die sie als Merkmale prädestinieren würde, wenn ihre Trennstärke nicht zu wünschen übrig ließe. Aus diesem Grund ist die Brauchbarkeit spontansprachlicher Merkmale mit einem Fragezeichen zu versehen.

n-Gramme

Für Bigramme und Trigramme schließlich gilt einerseits ähnliches wie bei den Vokabeln: Selbst die Gesamtzahl der am häufigsten vorkommenden n-Gramme führt, verteilt auf die einzelnen Sprecher, zu theoretischen Gesamtmengen in einem Bereich, der im Rahmen der Faustregel gerade noch oder bereits nicht mehr für einen χ^2-Test tauglich erscheint.

Obwohl auch die Variationskoeffizienten recht niedrig sind, werden im Hinblick auf eine sehr spezielle Stildefinition[34] Bigramme dennoch als Merkmale verwendet: Die Motivation dazu ist darin zu sehen, daß die endgültige Bewertung der Partitionen und Merkmalssätze weiter unten in 4.3 mit der Perplexität in einem stochastischen Sprachmodell aus n-Grammen vorgenommen werden wird; eben die Klassifikation der Sprecher nach n-Grammen sollte eine Spezialisierung bewirken, die das dazugehörige Modell für in dieser Hinsicht ähnliche Sprecher geeigneter macht[35].

Allerdings verbietet sich die Verwendung von Trigrammen wegen der zu geringen Häufigkeiten[36]. Auch bei Bigrammen ist nur die relative Häufigkeit verwendbar, während die turn- und satzbezogenen Werte zu gering ausfallen.

2.3 Erstellung von Merkmalssätzen

2.3.1 Vorüberlegungen

Grundsätzlich ist die Auswahl der später in die Clusteranalyse eingehenden Variablen in Merkmalssätzen eine nichttriviale Aufgabe; allgemeine Aussagen wie

> „Die Auswahl erfolgt jedesmal problembezogen. Es sollten jedesmal alle Merkmalsbereiche im gleichen Umfang durch die Variablenliste abgedeckt werden.“[37]

sind zwar durchaus einleuchtend, helfen aber in der Praxis nicht wesentlich weiter. Hier wird man um die heuristische Strategie *trial and error* nicht ganz herumkommen.

In der Literatur finden sich zur Verwendung von Merkmalssätzen mit statistischen Abhängigkeiten diametral entgegengesetzte Ansichten (vgl. [40], S. 34. f.): Einerseits wird auf die „clusterbildende Wirkung“ solcher Merkmale hingewiesen, die gelegentlich hochkorrelierte

[34] nämlich daß sich Stil in n-Grammen manifestiert, vgl. die Erläuterungen in 4.1

[35] Wir können Bigramme als Merkmale „außer Konkurrenz“ ansehen, deren Eignung — oder Nichteignung — sich erst am Schluß bei der Bewertung in 4.3 herausstellen wird.

[36] Die zugegebenermaßen etwas naive Herangehensweise an n-Gramme als mögliche Merkmale unter Ignoranz des in 4.1 angesprochenen Parameterproblems wird also für Trigramme sofort bestraft; für die Bigramme ist die Situation zum gegenwärtigen Zeitpunkt der Ausführungen noch nicht so klar.

[37] [40], S. 33 f.

Merkmale sogar erwünscht sein läßt, andererseits benötigt das Varianzkriterium[38] im wesentlichen unkorrelierte Merkmale mit gleicher Streuung (vgl. [6], S. 184); letzteres ist durch die Standardisierung erreichbar, und ersteres duch eine entsprechende Merkmalauswahl. — Unstreitig und damit der Hauptpunkt ist wohl die implizite stärkere Gewichtung eines Merkmalskomplexes durch die Auswahl mehrerer korrelierter Variablen aus seinem Bereich im gleichen Merkmalssatz.

Aus diesen Überlegungen heraus wurden in der vorliegenden Arbeit sowohl Merkmalssätze ohne korrelierte Merkmale als auch solche mit hochkorrelierten Merkmalen gebildet.

In inhaltlicher Hinsicht kann durch die quantitative Stilistik keiner der vorgenannten Bereiche a priori ausgeschlossen werden, sofern nur quantifizierbare Merkmale dazu ermittelt werden können, und eine Gewichtung ist sicherlich weitgehend Ansichtssache. Man kann so weit gehen zu fordern, daß persönlicher Stil sich in allen im Text auffindbaren Phänomenen ausdrücken muß, oder die andere Extremposition einnehmen und die Auffassung vertreten, Stil sei lediglich eine Frage der Verteilung von n-Grammen[39]: Lediglich der Einsatz in der Praxis der Sprachverarbeitung, und nicht die Konformität zu einer theoretisch noch so fundierten Stildefintion wird letztlich den Wert quantitativer Stilmerkmale zur Klassifikation von Sprechstilen erkennen lassen.

Eine weitere wichtige und nicht allgemeingültig zu beantwortende Frage betrifft die Zahl der in einem Merkmalssatz zusammenzufassenden Variablen. Auch wenn Bolz in [7], S. 215 als „Faustregel“ für eine Textgruppendifferenzierung einen Bereich zwischen 4 und 9 Merkmalen angibt, weil eine größere Zahl den statistischen Fehler vergößert und eine kleinere nicht ausreichend differenziert, so wird doch dieser Bereich in der vorliegenden Arbeit auf 2 bis 10 Merkmale ausgedehnt: 2 – 3 Merkmale ermögliche anschauliche Visualisierung und visuelle Kontrolle der Klassifikation, und 10 Merkmale wurden lediglich im weiter unten erläuterten Merkmalssatz *Top10Bigr* verwendet. Wenn man möglichst unkorrelierte Variablen verwenden möchte, so stellen allerdings bereits 4 bis 5 Merkmale die Obergrenze bei den in dieser Arbeit untersuchten Daten dar.

Also ergibt sich als Vorgehen eine pragmatische Strategie der schrittweisen Erweiterung: Von bereichsinternen zu bereichsübergreifenden, von unkorrelierten zu hochkorrelierten, von kleinen zu umfangreicheren Merkmalssätzen.

2.3.2 Wort-, Satz- und Turnlängen

Ein gerade für gesprochene, dialogische Sprache interessanter Merkmalssatz ist die Kombination aus Wort-, Satz- und Turnlängen (Turns treten ja nur bei Dialogen auf). Diese Zusammenstellung textunterteilender Merkmale unterschiedlicher Granularität zeigt sehr niedrige Korrelationen der einzelnen Variablen (Tabelle 2.1).

Allerdings weisen ADJ-Wortlängen, wie in 2.2.3 ausgeführt, eine höhere Variabilität auf als Wortlängen allgemein und eignen sich daher besser für einen Verwendung in einem Merkmalssatz. Zweiseitige Kolmogorov-Smirnov-Tests der Merkmale untereinander ergeben zwar zwischen den Satz- und Turnlängen keine hoch signifikanten Abweichungen, wohl aber zu den Wortlängen (s. Tabelle 2.2). Diese Überlegungen konstituieren den Merkmalssatz *WortSatzIII*.

[38] das Gütemaß der Wahl beim Clusteranalyseverfahren *k-means*

[39] die jedoch bei einem Linguisten oder gar einem Literaturwissenschaftler nur ein Kofschütteln hervorrufen dürfte

Tabelle 2.1: Korrelationen der Merkmale in *WortSatzII/III/IV/V*

	MADJWL	ADJWLE	AUSSPR/S	WORTSATZ	SATZTURN
MADJWL	1.000				
ADJWLE	0.480	1.000			
AUSSPR/S	0.124	0.142	1.000		
WORTSATZ	0.062	0.252	0.036	1.000	
SATZTURN	0.097	0.282	0.241	0.012	1.000

Tabelle 2.2: KOLMOGOROV-SMIRNOV-Werte der Merkmale in *WortSatzII/III/IV/V*

	MADJWL	ADJWLE	AUSSPR/S	WORTSATZ	SATZTURN
MADJWL	.				
ADJWLE	0.001	.			
AUSSPR/S	0.000	0.000	.		
WORTSATZ	0.002	0.000	0.000	.	
SATZTURN	0.022	0.000	0.001	0.294	.

Ersetzen wir in *WortSatzIII* die mittleren ADJ-Wortlängen durch die ADJ-Wortlängen-Entropie, eine Variable mit deutlich rechtssteiler Verteilung, so erhalten wir den neuen Merkmalssatz *WortSatzIV*. Die ADJ-Wortlängen-Entropie ist zwar eine etwas stärker mit den Satz- und Turnlängen korreliert, zeigt aber andererseits beim 2-Stichproben-KOLMOGOROV-SMIRNOV-Test noch höhere Signifikanz. Sie eignet sich daher besser als die ADJ-Wortlängen zur Verwendung im Merkmalssatz.

Wenn wir die Histogramme und Konfidenzellipsen der einzelnen Merkmalspaare in Abb. 2.5 betrachten, so lassen sich Form der Ellipsen anschaulich die Korrelationen zwischen den einzelnen Merkmalen erkennen: Verlaufen sie entlang der Diagonalen des ersten Quadranten, so ist die Korrelationhoch. Wie man sieht, ist dies jedoch bei den dargestellten Merkmalen nicht der Fall. Darüber hinaus erkennt man die unterschiedlichen Steilheiten der Histogramme und bekommt durch die XY-Darstellungen der einzelnen Merkmalspaare einen ersten Eindruck von der Vertilung der Datenpunkte.

WortSatzIV läßt sich in zwei Richtungen modifizieren: Zum einen kann man die Überlegung anstellen, zwecks schnellerer Klassifikation auf die relativ komplexe Berechnung der ADJ-Wortlängen-Entropie zu verzichten, weil diese als Diversitätsmaß einer Verteilung erst ab einer gewissen Datenmenge aussagekräftige Werte ergibt. Dieser Ansatz führt auf den nur aus Satz- und Turnlängen bestehenden Merkmalssatz *WortSatzIIa*.

Andererseits liegt der Ansatz nahe, *WortSatzIV* noch um ein spontansprachliches Merkmal mit hoher Variabilität zu erweitern und vielleicht durch die zusätzliche, domänenspezifische Information zu verbessern; aus Gründen der Korrelation mit der Turnlänge empfiehlt sich dafür eine satzbezogene Variable. Die Entscheidung fiel auf die Aussprachevarianten/Satz, weil sie die größten Unregelmäßigkeiten im Histogramm zeigen (s. Abb. 2.4): Die Idee dabei war, daß die relativ große Zahl der Sprecher ohne Aussprachevarianten möglicherweise eine eigene Klasse bilden könnte.

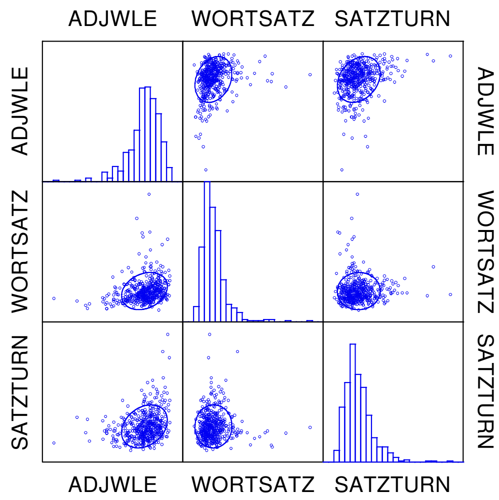

Abbildung 2.5: Histogramme von Wort-und Satzlängen

2.3.3 Bereichsübergreifende Merkmalssätze

Den Merkmalssätzen, die in diesem Abschnitt definiert werden, liegt sämtlich die Idee der Repräsentation aller Bereiche zugrunde, die in 2.3.1 zitiert wurde. Hinzu kommt das Bestreben, Variablen mit unterschiedlichen Basisgrößen mit einzubeziehen, so daß sich hier Satz- und Turnwerte und auch relative Häufigkeiten finden. Um jedoch die Zahl der Merkmale nicht zu hoch werden zu lassen, waren jeweils einige Zugeständnisse an dieses Ziel nötig — mit Ausnahme des letzten vorgestellten Merkmalssatzes *gesMerkmVII*.

gesMerkmVa, *gesMerkmVb* und *gesMerkmVc*

Die ADJ-Wortlängen-Entropie ADJWLE bietet sich neben der Verwendung bei den Satz- und Turnlängen auch als Wortlängenmerkmal an, mit dem wir einen neuen Merkmalssatz eröffnen. Will man eine Variable aus dem Bereich der Kategorien mit aufnehmen, so eignen sich bei den turnbezogenen Werten V/T (Verben/Turn) und bei den relativen Häufigkeiten ADV.

Die Merkmale weisen untereinander niedrige Korrelationen von 0,2 auf, bis auf den etwas höheren Wert von 0,3 zwischen V/T und ADV; dafür sind die Kolmogorov-Smirnov-Werte sehr niedrig, die Verteilungen also durchaus voneinander verschieden (s. die summarische Korrelationsmatrix in Tabelle 2.3 und die Kolmogorov-Smirnov-Matrix in Tabelle 2.4).

Möchte man darüber hinaus auch spontansprachliche Merkmale integrieren, so zeigt sich, daß eine Erweiterung von *gesMerkmVa* stets mit hohen Korrelationen mit V/T verbunden ist. Deshalb ist ein Austausch dieses Merkmals durch eine spontansprachlichen Variable angebracht. Die Merkmalssätze *gesMerkmVb* und *gesMerkmVc* sind solche Abwandlungen von *gesMerkmVa*, in denen V/T durch VERBAL/T (verbale Verzögerungen/Turn) bzw. NON-

Tabelle 2.3: Korrelationen der Merkmale in *gesMerkmVa/b/c*

	NONVER/T	VERBAL/T	V/T	ADV	ADJWLE
NONVER/T	1.000				
VERBAL/T	0.669	1.000			
V/T	0.670	0.678	1.000		
ADV	0.263	0.212	0.252	1.000	
ADJWLE	0.261	0.263	0.340	0.056	1.000

Tabelle 2.4: Kolmogorov-Smirnov-Werte der Merkmale in *gesMerkmVa/b/c*

	NONVER/T	VERBAL/T	V/T	ADV	ADJWLE
NONVER/T	.				
VERBAL/T	0.420	.			
V/T	0.572	0.148	.		
ADV	0.001	0.000	0.001	.	
ADJWLE	0.000	0.000	0.000	0.002	.

VER/T (nonverbale Verzögerungen/Turn) ersetzt ist.

gesMerkmVII

Schließlich wurde noch ein umfassender Merkmalssatz aus neun verschiedenen Variablen gebildet. Dabei wurden sowohl Wortlängenmerkmale als auch kategoriale sowie spontansprachliche Merkmale herangezogen.

Von den ersteren wurden neben der ADJ-Wortlängen-Entropie die Turnwerte bei der Anzahl mittlerer und langer Wörter verwendet. Die anderen Merkmale sind ebenfalls turnbezogen; es handelt sich um die beiden Verzögerungen, die Aussprachevarianten und die Füllwörter. Alle Turnwerte weisen eine recht hohe Korrelation auf, was bereits in der Ähnlichkeit ihrer Histogramme dutlich wird (in Abb. 2.6, S. 29 zeigt von den spontansprachlichen Merkmalen nur das Histogramm der Aussprachevarianten ein abweichendes Aussehen).

Die verstärkte Berücksichtigung des spontansprachlichen Bereichs soll jedoch trotz der hohen Korrelationen zwischen den Merkmalen die sprachlichen Spezifika des VM-Korpus repräsentieren, entsprechend der in 2.3.1 angeführten Hoffnung auf die clusterbildende Eigenschaft hochkorrelierter Merkmale.

Der Merkmalssatz *relEnde*

Von den relativen Häufigkeiten der Kategorien am Satzende wurde bereits gesagt, daß sie einen akzeptablen Variationskoeffizienten aufweisen und daß ihrer Untersuchung eine gewisse linguistische Motivation zugrundeliegt. Die Idee, sie in einem Merkmalssatz zusammenzufassen, liegt nahe, zumal sie fast ausschließlich niedrige Korrelationen aufweisen. Zwei Kategorien, nämlich V und ADJ, wurden ausgewählt; zur Begründung müssen wir an dieser Stelle einen

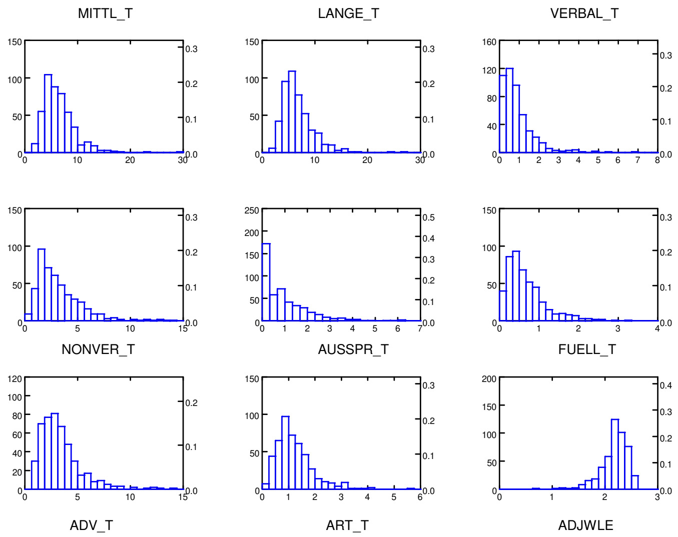

Abbildung 2.6: Histogramme von *gesMerkmVII*

kleinen Vorgriff auf die Klassifikation machen.

Bei der Betrachtung von Abb. 2.7 fällt die ausgeprägt durch Datenpunkte besetzte Diagonale auf, zu deren beiden Seiten sich die restlichen Punkte gruppieren. Ohne der linguistischen Interpretation vorzugreifen, läßt sich doch sagen, daß eine ganze Reihe Sprecher offensichtlich jeweils die gleiche Zahl von V und ADJ am Satzende haben. Diese visuelle Klassifikation , die im übrigen auch in der Literatur durchaus ihren Stellenwert hat (vgl. [28], S. 63) — sie wurde zur Verdeutlichung bereits in die Grafik eingetragen — wird im weiteren Verlauf mit mathematischen Verfahren überprüft und verglichen werden.

2.3.4 Zusammenfassung

Fassen wir die ermittelten Merkmalssätze noch einmal in einer Tabelle zusammen und bezeichnen sie mit Kürzeln[40], die sie im folgenden identifizieren sollen (Tabelle 2.5, S. 30).

Die Kürzel sind neben ihrer mnemotechnischen Funktion auch dazu geeignet, die Aufstellung der Merkmalssätze zu strukturieren: Wir haben es mit zwei größeren Gruppen zu tun, von denen eine Merkmalssätze enthält, die aus Wort-, Satz- und Turnlängenmerkmalen zusammengesetzt sind (*WortSatz...*), während die andere bereichsübergreifende, v. a. spontansprachliche Merkmale enthält (*gesMerkm...*). Dazu kommt als Sonderfall der Merkmalssatz *relEnde* aus Kategorien am Satzende, der mehr experimentellen Charakter hat. — Schließlich wurden entsprechend den obigen Überlegungen die relativen Häufigkeiten der 10 häufigsten Bigramme in dem umfangreichen Merkmalssatz *Top10Bigr* zusammengefaßt.

Mit diesen Merkmalssätzen werden wir im nächsten Kapitel Clusteranalysen durchführen.

[40] die Namensgebung der Kürzel hat ihren Ursprung in der Geschichte der Bearbeitung der Merkmalssätze

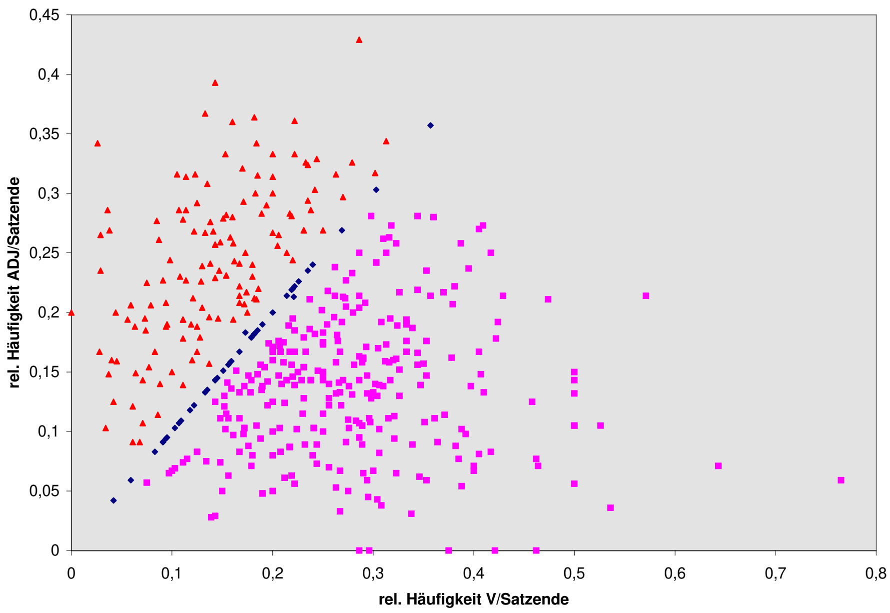

Abbildung 2.7: XY-Plot von *relEnde* mit eingetragener visueller Klassifikation

Tabelle 2.5: Zusammmenstellung der Merkmalssätze

Kürzel	Merkmale	Beschreibung
gesMerkmVa	3	Verben/Turn, rel. ADV-Häufigkeit, ADJ-Wortlängen-Entropie
gesMerkmVb	3	verbale Verz./Turn, rel. AP-Häufigkeit, ADJ-Wortlängen-Entropie
gesMerkmVc	3	nonverbale Verz./Turn, rel. AP-Häufigkeit, ADJ-Wortlängen-Entropie
gesMerkmVII	9	Turnwerte von mittl./langen Wörtern, (non)verbale Verz., Ausspr., Füllw., ART, ADV, ADJ-Wortlängen-Entropie
Top10Bigr	10	rel. Häufigkeit der 10 häufigsten Bigramme
relEnde	2	rel. V/Satzende, rel. ADJ/Satzende
WortSatzIIa	2	Wörter/Satz, Sätze/Turn
WortSatzIII	3	Wörter/Satz, Sätze/Turn, mittl ADJ-Wortlänge
WortSatzIV	3	Wörter/Satz, Sätze/Turn, ADJ-Wortlängen-Entropie
WortSatzVa	4	Wörter/Satz, Sätze/Turn, ADJ-Wortlängen-Entropie, Aussprachevarianten/Satz

Kapitel 3

Ermittlung der Klassenstrukturen

Beginnen wir nun wie angekündigt damit, die im vorigen Kapitel aufgstellten Merkmalssätze auf Klassenstrukturen zu untersuchen. Es wurde schon angedeutet, daß dies mit Hilfe clusteranalytischer Verfahren geschehen wird[1]. Wir gehen im folgenden davon aus, daß die zu analysierenden Daten in einer Matrix vorliegen, der sog. Datenmatrix X_{ij}, deren Zeilen die Beobachtungen und deren Spalten die Merkmale darstellen.

3.1 Clusteranalyse

Am Anfang dieses Abschnittes soll eine Aussage in [40], S. 44 stehen.

> „Wenn in den Daten eine ausgeprägte Clusterstruktur vorhanden ist, so wird diese von verschiedenen Verfahren gefunden werden.“

Dieses Zitat ist als Aufforderung an den für statistische Details wenig begeisterten Leser zu sehen, die folgenden Ausführungen zu überspringen und sich gleich der Darstellung der Ergebnisse in 3.2 zuzuwenden.

3.1.1 Abstandsmaße und Standardisierung

Bevor überhaupt eine Analyse durchgeführt werden kann, sind für die Daten geeignete Abstandsmaße zu definieren. Das gebräuchlichste Maß für quantitative Daten — und nur mit solchen haben wir es im Rahmen der quantitativen Stilistik zu tun — ist der EUKLIDische Abstand, nicht zuletzt wegen seiner direkten geometrischen Interpretation.

Dieses Distanzmaß hat aber den Nachteil, nicht skaleninvariant zu sein[2], weshalb die Daten vor seiner Verwendung zu standardisieren sind[3]; die Methode der Wahl dazu ist die z-Transformation (s. 2.2.2), durch deren Anwendung eine Variable den Mittelwert 0 und Standardabweichung 1 erhält.

[1] Die Berechnungen und grafischen Auswertungen zu den Clusteranalysen in dieser Arbeit wurden mit den Programmen *ClusCorr* für *MS Excel* und *Systat* für Windows auf einem PC vorgenommen.

[2] ein einfaches Beispiel dazu findet sich z.B. in [10], S. 21 f.

[3] Die Alternative wäre das Ausweichen auf ein sklaeninvariantes Merkmal wie den MAHALANOBIS-ABSTAND. Nachdem aber die Standardisierung zur Verfügung steht, ist nicht einzusehen warum man nicht beim EUKLIDischen Abstand bleiben sollte.

Handelt es bei den Elementen der Datenmatrix um Häufigkeiten, inbesondere um Prozentwerte, so empfiehlt MUCHA ([28], S. 79) den χ^2–Abstand[4], der in dieser Arbeit allerdings nur bei einem entsprechenden Merkmalssatz verwendet wurde.

Eng zusammen mit der Wahl des Distanzmaßes hängt die Frage nach der Gewichtung der einzelnen Variablen; man kann u. a. mit den reziproken Werten von Standardabweichung, Wertebereich oder Mittelwert gewichten und so die Standardisierung ganz oder teilweise ersetzen. In der vorliegenden Arbeit wurden die Merkmalssätze aus relativen Häufigkeiten mit dem ersten der angesprochenen Werte multipliziert, um die teilweise doch recht unterschiedlichen Standardabweichungen auszugleichen.

Auch Beobachtungen lassen sich gewichten, was insbesondere dazu benutzt werden kann, von vorneherein identifizierte Ausreißer aus der weiteren Analyse auszuschließen, indem ihnen das Gewicht 0 zugeteilt wird[5].

Zur Datenreduktion wie auch zur visuellen Inspektion dient die sog. Hauptkomponentenanalyse, im Prinzip eine Projektion des Merkmalsraums auf eine kleinere Anzahl von Komponenten mit Hilfe der Eigenwerte der empirischen Kovarianzmatrix (vgl. [6], S. 238), die nach ihrem Beitrag zur Varianz der Punktwolken geordnet sind (vgl. [28], S. 63 ff.) und keine Korrelation mehr aufweisen (vgl. [40], S. 102). Weiterhin ist die Überprüfung und Bewertung der Klassifikation durch eine graphische Darstellung der ersten beiden Hauptkomponenten möglich (vgl. [40], S. 46). Die Hauptkomponentenanalyse kann somit am Ende einer Clusteranalyse stehen, wie auch Ausgangspunkt für weitere Analysen sein.

Nach der Wahl eines geeigneten Distanzmaßes faßt man die Abstände der einzelnen Beobachtungen voneinander in der Distanzmatrix zusammen, auf der dann die verschiedenen Verfahren durchgeführt werden können, ohne jedesmal die Distanzen neu berechnen zu müssen.

3.1.2 Auswahl der Verfahren und Klassenzahlen

Eine notwendige Vorüberlegung betrifft die Frage nach der maximalen sinnvoll anzunehmenden Klassenzahl, die für diese Arbeit auf $6-7$ festgesetzt wurde: Bei höheren Klassenzahlen treten bei jeder Klassifikation einzelne fast leere Cluster auf ($3-4$ Sprecher sind keine Seltenheit), so daß die so entstehenden „Cluster" sehr heterogen werden und im Grunde nur noch „Ausreißer" enthalten.

Hat man die Datenmatrix durch eine geeignete Standardisierung vorberereitet und mit dem gewählten Abstandsmaß eine Distanzmatrix berechnet, so stellt sich die Frage nach einem passenden Clusterverfahren. Generell unterscheidet man bei der Clusteranalyse hierarchische und partitionierende Methoden.

Partitionierende Clusteranalyse

Partitionierende Verfahren zerlegen die Daten in eine fest vorgegebene Anzahl von Klassen, wobei leere Klassen prinzipiell zugelassen sind, aber natürlich nicht vorkommen sollten. Wenn die Klassenzahl nicht a priori bekannt ist, müssen die Verfahren für jede in Frage kommende

[4] Andererseits ist der χ^2–Test keinesfalls auf Prozentzahlen anzuwenden (vgl. [29], S. 119). Der sich aus diesen beiden Aussagen scheinbar ergebende Widerspruch läßt sich jedoch auflösen, wenn man bedenkt, daß des sich beim χ^2–Abstand um ein Distanzmaß und nicht um eine Teststatistik handelt, für die andere Anwendungsbedingungen gelten.

[5] Von dieser Möglichkeit wurde jedoch im Rahmen der vorliegenden Arbeit kein Gebrauch gemacht, weil die visuelle Inspektion der Daten entweder keine klaren Ausreißer oder — je nach Betrachtungsweise — zu viele Ausreißer erkennen ließ, um alle auszuschließen.

Klassenzahl einmal gesondert durchgeführt werden. Ein typisches Verfahren dieser Gruppe ist *k-means*, eine spezielle Art des iterativen Minimaldistanzverfahrens, bei dem das Intraklassen-Varianzkriterium V_K für die Datenmatrix X_{ij} optimiert wird ([28], S. 122).

$$V_K = \sum_{k=1}^{K} \sum_{i \in C(k)} u_i d_{ik}^2$$

mit den Variablengewichten u_i und dem Distanzwert d_{ik}, für den nach MUCHA der quadrierte EUKLIDische Abstand

$$d_{ik}^2 = \sum_{j=1}^{J} q_j (x_{ij} - \bar{x}_{kj})^2$$

den Standard darstellt (dabei sind die $\bar{x}_{kj}$ die Mittelwerte der Klasse k der Variablen in der Spalte j und q_j die Variablengewichte).

MUCHA sieht den quadrierten EUKLIDischen Abstand zwar als das Standardmaß an, aber natürlich sind statt dessen auch andere Distanzmaße denkbar.

Mit dem Kriterium V_K wird die Intraklassenvarianz minimiert und dabei die Interklassenvarianz maximiert mit dem Ziel, nach möglichst wenigen Iterationen homogene und separierte Cluster zu erhalten. Ein Nachteil dieses Verfahrens ist seine Reihenfolgeabhängigkeit in Bezug auf die Beobachtungen in der Datenmatrix, die zu völlig unbrauchbaren Ergebnissen bei klassenweise geordneter Datenmatrix führen kann — allerdings muß man diesem theoretischen Extremfall, der für eine unbehandelte, initiale Datenmatrix unwahrscheinlich ist, nicht unbedingt zu starke Bedeutung beimessen. Der Vorteil ist, daß *k-means* nur zulässige, d. h. nichtleere Cluster erzeugt.

Außerdem tendiert das Varianzkriterium zur Bildung kugelförmiger Cluster[6]; dies kann im Falle stark korrelierter Merkmale unerwünscht sein (vgl. [40], S. 167 f.). Als Abhilfe kann man entweder zu einem anderen Kriterium greifen (vgl. [6], S. 184 f.), eine Hauptkomponententransformation vornehmen ([40], S. 168) oder von vorneherein relativ unkorrelierte Variablen verwenden und diese standardisieren; der letzere Weg wurde in dieser Arbeit gewählt[7].

Zur Festlegung einer plausiblen Klassenzahl kann man bei *k-means* wie auch bei hierarchischen Verfahren die Intra- und InterklassenVarianzsummebilden und ihre Entwicklung mit steigender Klassenzahl verfolgen (wobei die erste Summe monoton fällt und die zweite monoton zunimmt). Man wird dann die Klassenanzahl auswählen, bei der sich die beiden Summen „deutlich“ (vgl. [6], S. 194) von den Werten der vorhergehenden und nur unwesentlich von denjenigen der nächsthöheren Klassenzahl unterscheiden. Natürlich sind diese qualitativen Angaben zu einem gewissen Grad der subjektiven Einschätzung unterworfen. Man wird gut daran tun, diese qualititativen Kriterien an grafischen Veranschaulichungen auszuwerten.

Ursprünglich wurde für die Untersuchungen dieser Arbeit auch das Austauschverfahren als weitere partitionierende Methode auf einem Teil der Trainingsmenge verwendet, aber wegen der mit *k-means* fast durchwegs übereinstimmenden Clustereinteilungen wurde die Anwendung dieses Verfahrens nicht auf die gesamte Trainingsstichprobe ausgeweitet[8].

---

[6] in Bezug auf die Distanzfunktion, bei EUKLIDischem Abstand handelt es sich tatsächlich um Hyperkugeln, die man jedoch durch geeignete Gewichtung der Variablen auch aus Ellipsoiden erzeugen kann

[7] einmal aus theoretischen Überlegungen der Flexibilität des Varianzkriteriums (vgl. [40], S. 166) und zum anderen aus rein praktischen Gründen der Verfügbarkeit in der verwendeten Software, die aber auch ein Indiz für die begründeteten Präferenzen der Autoren ist.

[8] vgl. [28], S. 125 zur Ähnlichkeit der beiden Verfahren

Hierarchische Clusteranalyse

Hierarchische Verfahren zeichnen sich dadurch aus, daß sie eine Klassenhierarchie aufbauen, ohne daß dabei eine feste Klassenzahl vorgegeben werden muß. Je nachdem, ob die jeweilige Methode dabei die gesamte Stichprobe als Initialklasse annimmt, die schrittweise zerlegt wird, oder anfangs die einzelnen Beobachtungen als „Trivialcluster“ ansieht, die im Verlauf des Algorithmus zusammenzufassen sind, spricht man von divisiven bzw. agglomerativen Verfahren[9].

Die verschiedenen Verfahren der hierarchisch-agglomerativen Clusteranalyse unterscheiden sich im wesentlichen durch die Kriterien zur Festlegung der Klassendistanz, während der grundlegende Ablauf eines Iterationsschritts gleich bleibt (vgl. [28], S. 96 f.): Es wird zunächst der minimale Abstand $d_{ll'}$ zweier verschieder Elemente in der Distanzmatrix gesucht, dann werden die beiden zugehörigen Klassen l und l' zu einer neuen Klasse k fusioniert und schließlich die Abstände $t_{kk'}$ von k zu allen anderen Klassen k' nach

$$t_{kk'} = \alpha_l d_{lk'} + \alpha'_l d_{l'k'}) + \beta d_{ll'} + \gamma |d_{lk'} - d_{l'k'}|$$

berechnet. Dabei sind α_l, α'_l, β und γ die Parameter zur Festlegung der Klassendistanz $t_{kk'}$, die die einzelnen Unterverfahren kennzeichnen.

Der Algorithmus bricht ab, wenn keine Fusionen mehr möglich sind, also nur noch eine Klasse übrig ist, die dann alle Beobachtungen enthält. Aus der Fusionsabfolge läßt sich eine Hierarchie der Daten ableiten, die dem Verfahren den Namen gegeben hat.

Beim Standardverfahren WARD *linking*, das von einem Modell des EUKLIDischen Vektorraums ausgeht, wird im Verlauf des Fusionsprozesses der anfänglich einelementigen „Klassen“ zu größeren Clustern in jedem Algorithmusschritt beim Übergang von $K-1$ auf K Klassen der Varianzzuwachs V_K

$$\Delta V_K = V_K - V_{K-1}$$

innerhalb aller Klassen minimiert. V_K ist wie bei den partitionierenden Verfahren definiert.

Die obigen Parameter zur Bestimmung der Klassendistanz ergeben sich mit den Gewichten u_i und u'_i der Beobachtungen i und i' bei WARD *linking* wie folgt:

$$\alpha_l = \frac{u_l + u'_k}{u_l + u'_l + u'_k} \quad , \quad \alpha'_l = \frac{u'_l + u'_k}{u_l + u'_l + u'_k} \quad , \quad \beta = \frac{-u'_k}{u_l + u'_l + u'_k} \quad , \quad \gamma = 0$$

WARD *linking* tendiert dazu, am Anfang des Algorithmus viele kleine Klassen in Regionen hoher Punktdichte zu erzeugen und große Klassen mit terminalen Klassen zu fusionieren; oft werden isoliert liegende Punkte in der Datenmatrix durch WARD *linking* nicht erkannt.

Was die Festlegung der Klassenzahl bei den hierarchischen Verfahren angeht, so gibt MUCHA multivariate und univariate Gütekriterien an, deren Berechnung aufwendig ist und die dann doch in einer graphischen Darstellung untersucht werden müssen, weil „Signifikanztests [...] nur bei a priori gegebener Klasseneinteilung erlaubt“[10] sind. — Wir bevorzugen in dieser Arbeit mit STAHL, der „für die Festlegung einer geeigneten Clusterzahl [...] das Dendrogramm als eines der wichtigsten Hilfsmittel“[11] ansieht, die Inspektion der grafisch veranschaulichten Fusionsniveaus sowie die bei hierarchischen Methoden üblichen Dendrogramme, und zwar in

[9] Für den Informatiker sind die Bezeichnungen *top-down* und *bottom-up* für diese Untergliederung unmittelbar einsichtig; sie tauchen aber meines Wissens in der Theorie der Clusteranalyse nicht auf.

[10] [28], S. 112

[11] [40], S. 163

der von *ClusCorr* zur Verfügung gestellten Variante des Dendrogramms in der ersten Hauptkomponentenebene.

Der grundsätzliche Nachteil hierarchischer Verfahren, daß einmal vorgenommene Klassenzuordnungen nicht mehr rückgängig gemacht werden können, ist auch bei WARD *linking* festzustellen. In dieser Möglichkeit liegt der Vorzug partitionierender Verfahren, die nach [28], S. 97 im Mittel den hierarchischen Verfahren überlegen sind[12].

Am Ende einer Clusteranalyse[13] steht gewönhnlich die Bewertung der Ergebnisse, die die Untersuchung clusterbezogener Mittelwerte und Varianzen sowie schließlich die inhaltliche Interpretation der Cluster umfaßt. Weil Clusterverfahren unabhängig von der tatsächlichen Existenz von Klassen stets ein Resultat liefern, stellt sich stets die Frage, ob eine etwaige „natürliche Klassenstruktur" durch die Analyse aufgedeckt werden konnte. Findet sich eine solche jedoch nicht, so muß man sich mit einer „überschaubare[n] und gut interpretierbare[n] Zerlegung der Objektmenge" ([40], S. 45) ausgehen und die klassischen allgemeinen Kriterien der Homogenität und Separiertheit von Clustern in den Hintergrund treten lassen.

3.2 Anwendung auf die Merkmalssätze

Nachdem aus der Theorie und von der Struktur der Variablen her nicht von vorneherein festzulegen war, ob eine hierarchische oder eine partitionierende Clusteranalyse für die in 2.3 aufgestellten Merkmalssätze zu bevorzugen ist, und um eine Vergleichsbasis für die später folgende Bewertung zu schaffen, wurden für jeden Merkmalssatz Clusteranalysen mit den Verfahren *k-means* und WARD *linking* durchgeführt. Grundsätzlich wurden die Variablen mit der z-Transformation standardisiert, um allen Variablen den gleichen Stellenwert einzuräumen und mit dem einfachen, allgemein gebräuchlichen, und leicht geometrisch interpretierbaren EUKLIDischen Abstand als Distanzmaß arbeiten zu können. Weder Merkmale noch einzelne Beobachtungen wurden bei der Clusteranalyse mit speziellen Gewichten versehen.

Im Falle von ausschließlich aus relativen Häufigkeiten bestehenden Merkmalssätzen wurde allerdings nicht der EUKLIDische, sondern der χ^2-Abstand als Distanzmaß verwendet. Da diese Variablen ohnehin in ihrem Wertebereich auf das Intervall $[0; 1]$ beschränkt sind, wurde auch keine Standardisierung vorgenommen, sondern die Merkmale bei der Clusteranalyse mit Gewichten versehen, nämlich ihrer inversen Standardabweichung.

Weil alle gesammelten Ergebnisse der Clusteranalysen in 4.3 bewertet werden, also keinerlei Partitionen für die Bewertung und Interpretation verloren gehen, wird das Hauptproblem der folgenden Abschnitte die Festlegung von geeigneten Klassenzahlen für den jeweiligen Merkmalssatz sein. In den Fällen, in denen entweder die hierarchische oder die partitionierende Analyse keine Präferenzen für eine bestimmte Klassenzahl erkennen ließ, wurde der Wert des jeweils anderen Verfahrens ausgewählt, um einen direkten Vergleich der Ergebnisse beider Verfahren zu ermöglichen.

Die Homogenität und Separiertheit der ermittelten Cluster kann für die partitionierende Clusteranalyse mit dem Verfahren *k-means* über die von *Systat* erzeugten Clusterprofile begutachtet werden, die clusterbezogene Mittelwerte und Standardabweichung en in den einzelnen Variablen veranschaulichen; bei der hierarchischen Methode WARD *linking* können die Cluster direkt über das mit *ClusCorr erzeugte* Hauptkomponenten-Plotdendrogramm auf Homogenität und Separiertheit überprüft werden. Diese Grafiken werden jedoch nicht durchgängig

[12] was, wie wir noch sehen werden, für die Daten des VM-Korpus nicht unbedingt zutrifft
[13] vgl. [40], S. 43 ff.

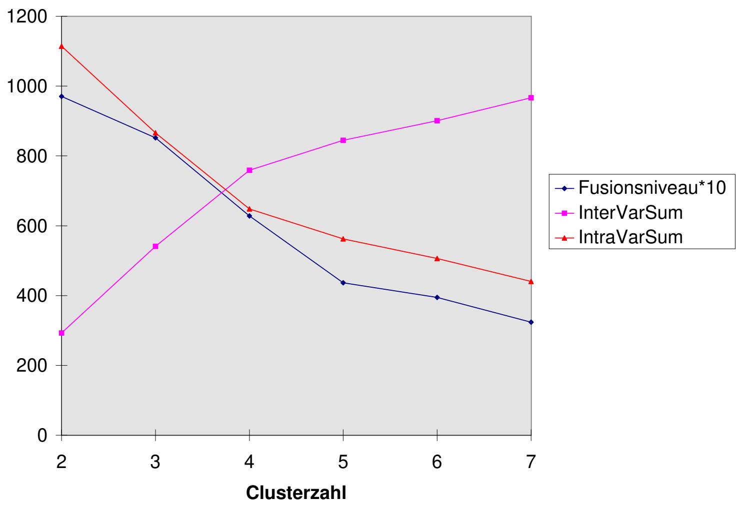

Abbildung 3.1: Varianzsummen und Fusionsniveaus von *WortSatzIII*

aufgeführt, einerseits, um der Bewertung mit einem anwendungsspezifischen Gütemaß und der Interpretation in den nächsten Kapiteln nicht vorzugreifen, und andererseits, um die Darstellung nicht unnötig in die Länge zu ziehen. Wir werden uns im Normalfall mit summarischen Aussagen begnügen.

3.2.1 Satz- und Turnlängen

Zunächst werden die clusteranalytischen Verfahren auf die eine der beiden in 2.5 aufgestellten Gruppen von Merkmalssätzen angewendet, nämlich die aus Wort-, Satz- und Turnlängenmerkmalen bestehenden Merkmalssätze.

Analyse von *WortSatzIII*

Die linguistische Anschaulichkeit und die Einfachheit der Merkmale läßt einen Beginn der Untersuchungen mit *WortSatzIII* gerechtfertigt erscheinen.

Führt man die Clusteranalyse mit dem Verfahren *k-means* und dem EUKLIDischen Abstand durch, so zeigen sich bei den Inter- und Intraklassen-Varianzsummen Sprünge beim Übergang von 3 auf 4 Klassen und eine relativ geringe Änderung beim folgenden Übergang auf 5 Klassen (Abb. 3.1). Dies spricht für die Klassenzahl 4, bei deren Annahme sich eine Partitionierung der Trainingsmenge in 81:167:186:36 Sprecher ergibt.

Nach Anwendung des Verfahrens WARD *linking* zeigt sich jedoch im Graphen der Fusionsniveaus (s. ebenfalls Abb. 3.1) ein deutlicher Knick nach unten bei der Klassenzahl 5, was ein Indiz für die Annahme dieser Klassenzahl ist. Die Partitionierung ergibt sich in diesem Falle zu 220:143:90:4:13, d. h. „Klasse“ 4 enthält augenscheinlich nur Datenausreißer, die aber ausreichen, um den erwähnten Knick bei den Fusionsniveaus hervorzurufen.

Es liegt hier also der Fall vor, daß die hierarchische und die partitionierende Clusteranalyse verschiedene Klassenzahlen favorisieren; eine entgültige Entscheidung muß der Bewertung in 4.3 vorbehalten bleiben.

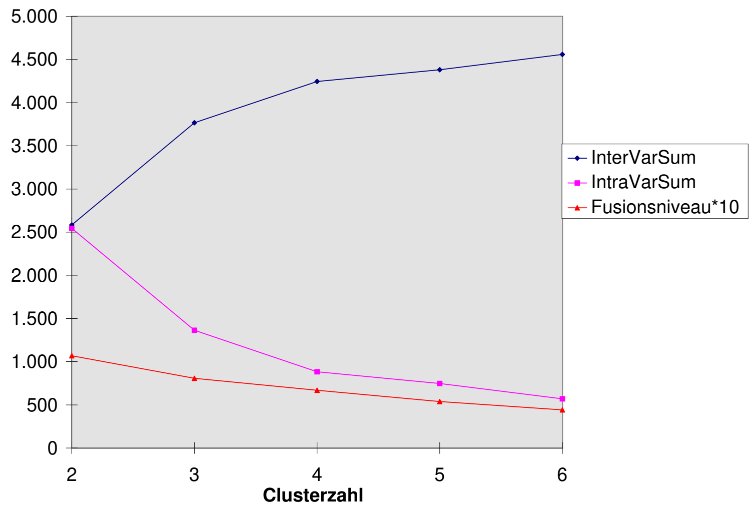

Abbildung 3.2: Varianzsummen und Fusionsniveaus von *WortSatzIV*

Analyse von *WortSatzIV*

Ein erster Versuch, nach Abschluß der hierarchischen Clusteranalyse eine plausible Klassenzahl für *WortSatzIV* zu ermitteln, führt über die Inspektion der Fusionsniveaus in Abb. 3.2; zwischen 3 und 5 Klassen verläuft der Graph jedoch fast als Gerade, so daß WARD *linking* keine von diesen Klassenzahlen präferiert.

Deshalb wurde eine Analyse der Hauptkomponenten vorgenommen und ein Dendrogramm in der ersten Hauptkomponentenenebene aufgestellt (s. Abb. 3.3). Um eine sinnvolle Klassenzahl zu ermitteln, betrachten wir die Verteilung der Datenpunkte in Abb. 3.3, speziell die markierten Cluster 3 und 4: Würde man lediglich 3 Cluster annehmen, so müßten, wie aus den Verzweigungen des über die Hauptkomponentenebene gelegten Dendrogramms ersichtlich ist, die Cluster 3 und 4 zu einem augenscheinlich sehr heterogenen Cluster zusammengefaßt werden; man tut also gut daran, bei einer Klassenzahl von mindestens 4 zu bleiben. Andererseits erscheint eine Aufteilung in mehr als 4 Cluster nicht sinnvoll (die Struktur der Daten läßt keine weiteren Verfeinerungsmöglichkeiten erkennen).

Im übrigen ist in Abb. 3.3 bereits ein für die Daten der vorliegenden Arbeit charakteristisches Phänomen zu beobachten: Eine wirkliche "natürliche Klassenstruktur" im Sinne STAHLS ist nicht zu erkennen, so daß bereits zu diesem Zeitpunkt abzusehen ist, daß der anwendungsspezifischen Bewertung und Interpretation erhöhte Bedeutung beizumessen sein wird.

Die nach der Clusteranalyse mit *k-means* und dem Distanzmaß EUKLIDischer Abstand entstehenden Inter- und Intraklassen-Varianzsummen sind in Abb. 3.2 eingetragen; man erkennt dort Sprünge beim Übergang von 3 auf 4 Klassen und eine relativ geringe Änderung beim folgenden Übergang auf 5 Klassen, so daß für *k-means* eine Klassenzahl 4 sinnvoll erscheint.

All dies zusammengenommen, ist die Klassenzahl 4 für *WortSatzIV* als die plausibelste anzusehen. Die Partitionierung ergibt sich damit für WARD *linking* in 272:62:103:33 und für

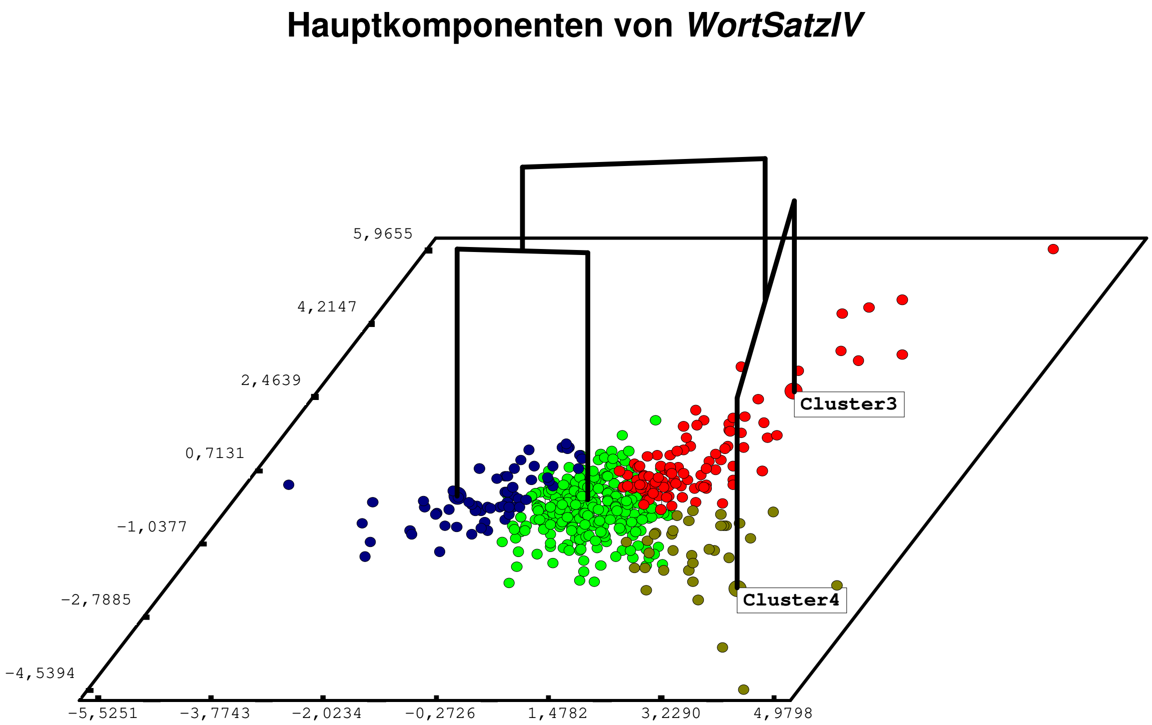

Abbildung 3.3: Hauptkomponenten-Dendrogramm von *WortSatzIV*

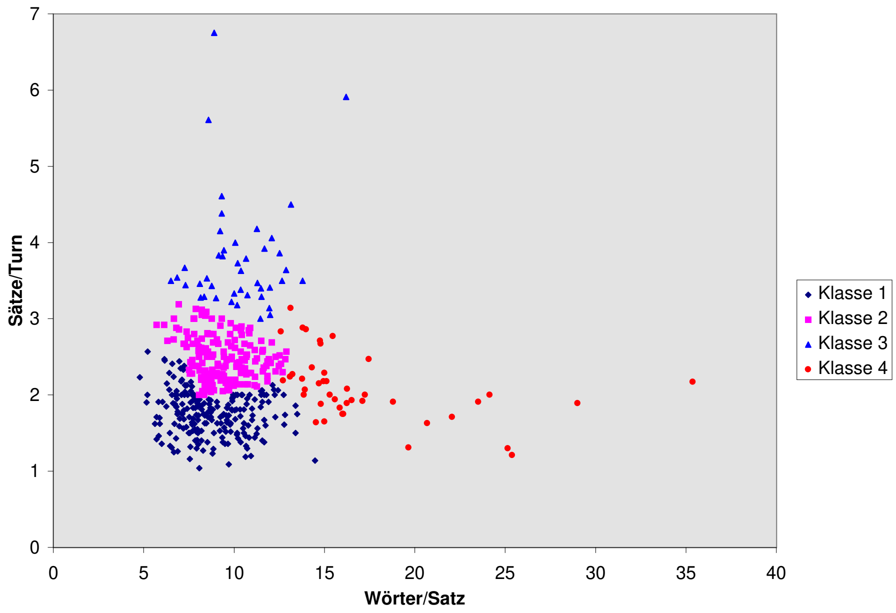

Abbildung 3.4: XY-Clusterplot von *WortSatzIIa* mit WARD *linking*

k-means in 78:36:87:269 Sprecher.

Analyse von *WortSatzIIa* und *WortSatzVa*

Für *WortSatzIIa* stand eine andere Klassenzahl als die bei *WortSatzIV* ermittelte nicht zur Debatte, da dieser Merkmalssatz als Vereinfachung von *WortSatzIV* definiert wurde und somit dessen Klassenstruktur möglichst verlustfrei nachbilden sollte.

Weil *WortSatzIIa* nur zwei Merkmale enthält, kann auf eine Analyse der Hauptkomponenten verzichtet und eine direkte grafische Darstellung angegeben werden (Abb. 3.4). Man erkennt, daß die Klassen 3 und 4 nicht besonders gut separiert sind, was auf das Fehlen der dritten Variable ADJ-Wortlängen-Entropie zurückzuführen ist, die diese beiden Cluster trennt (s. Abb. 5.1, S. 63); dies ist eben der Preis, der für die geringere Komplexität des Merkmalssatzes *WortSatzIIa* gegenüber *WortSatzIV* zu zahlen ist. — Die Partitionierung ist für *k-means* durch die Klassengrößen 81:167:186:36 und für WARD *linking* durch 218:167:43:42 beschrieben.

Wenn *WortSatzIIa* eine Reduktion von *WortSatzIV* darstellt, so war *WortSatzVa* als Erweiterung dieses Merkmalssatzes konzipiert mit dem Ziel einer Verbesserung der Klassifikation durch Hinzuziehen eines spontansprachlichen Merkmals.

Das Verfahren *k-means* favorisiert aber nicht die für *WortSatzIV* ermittelte Klassenzahl 4, sondern 5 Klassen: Wie man Abb. 3.5 entnimmt, ist die Interklassen-Varianzsumme für 4 Klassen sogar noch kleiner als die Intraklassen-Varianzsumme. Die Partitionierung ergibt sich in diesem Falle zu 207:109:55:31:68.

Andererseits zeigt sich an den ebenfalls in Abb. 3.5 eigetragenen Fusionsniveaus der hierarchischen Clusteranalyse, daß das Verfahren WARD *linking* eher für 4 Klassen spricht, wodurch

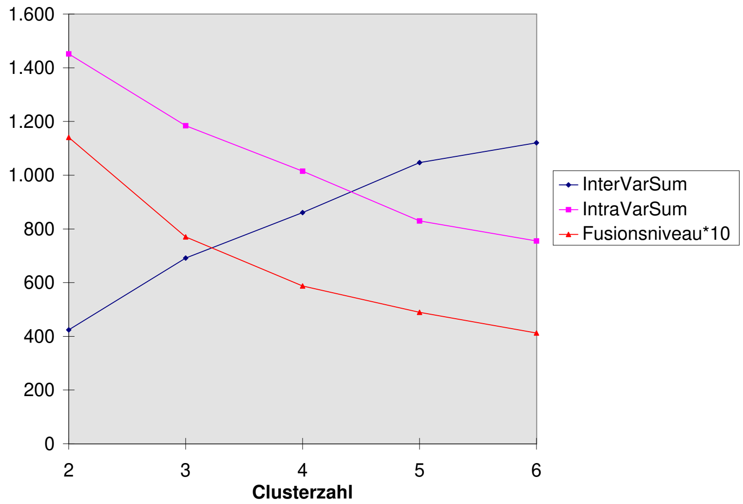

Abbildung 3.5: Varianzsummen und Fusionsniveaus von *WortSatzVa*

die Trainingsmenge in Klassen der Größen 115:171:92:92 partitioniert wird. Die Klassenzahl ist also hier wie im Falle von *WortSatzIII* unklar, und wie dort wird erst die abschließende Bewertung in Abschnitt 4.3 zeigen, welche Klassenzahl die besseren Ergebnisse bringt und damit als die korrekte zu gelten hat.

3.2.2 Bereichsübergreifende Merkmalssätze

Wenn jetzt die zweite größere Gruppe der in 2.5 aufgestellten bereichsübergreifenden Merkmalssätze betrachtet wird, so kann die Darstellung etwas abgekürzt werden, da das Prinzip im vorangegangenen Abschnitt deutlich geworden sein sollte und sich sonst die Wiederholungen häufen würden.

Analyse von *gesMerkmVa*, *gesMerkmVb* und *gesMerkmVc*

Bei *gesMerkmVa* stellt sich nach Anwendung der hierarchischen Clusteranalyse wieder die Frage nach der Klassenzahl.

Für die partitionierende Analyse können wir sie mit Hilfe von Abb 3.6 beantworten: Die dort eingetragene Interklassen-Varianzsumme wird erst ab 4 Klassen größer als die Intraklassen-Varianzsumme, danach aber auch nicht mehr „deutlich“ kleiner, so daß für *k-means* von 4 Klassen auszugehen ist. Die ebenfalls eingetragenen Fusionsniveaus sprechen für keine spezielle Klassenzahl bei dem Verfahren Ward *linking*; also wurde für die hierarchische Clusteranalyse derselbe Wert wie für die partitionierende angenommen. Damit erhalten wir eine Partitionierung in 89:136:89:156 Sprecher für Ward *linking* und 94:61:164:151 Sprecher für *k-means*.

Beim Vergleich von Abb. 3.6 mit der Darstellung der entsprechende Werte für *gesMerkmVb* in Abb. 3.7 fallen die deutlichen Ähnlichkeiten in den Ergebnissen beider Merkmalssätze ins

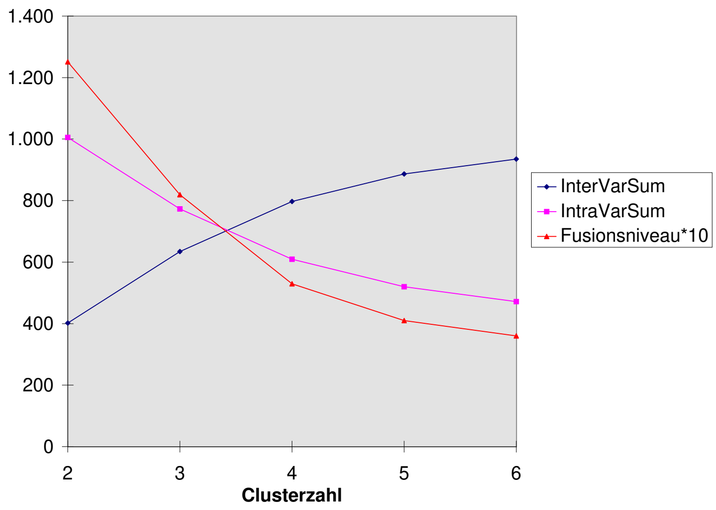

Abbildung 3.6: Varianzsummen und Fusionsniveaus von *gesMerkmVa*

Auge; man erkennt eine (allerdings nicht sehr ausgeprägte) Präferenz beider Verfahren für 4 Klassen[14].

Der Graph der Fusionsniveaus von *gesMerkmVb* in Abb. 3.7 verläuft von der Klassenzahl 2 bis zur Zahl 4 gerade und zeigt dann einen nach unten gerichteten Knick, woraus sich eindeutig 4 Klassen auch für dieses Verfahren als plausibel ergeben.

Man wird bei *gesMerkmVb* also ebenfalls von 4 Klassen auszugehen haben. Die Partitionierung für diese Klassenzahl ergibt sich für *k-means* zu 174:42:184:70 und für WARD *linking* zu 134:126:107:103.

Wie sich schon aus der Definition von *gesMerkmVc* ergibt, ähnelt dieser Merkmalssatz in mancher Hinsicht *gesMerkmVb*. In Abb. 3.8 zeigt sich jedoch, daß die Klassenzahl für WARD *linking* nicht auf 4 anzusetzen ist, weil der Graph der Fusionsniveaus dort gerade verläuft. Den leichten Knick bei der Clusterzahl 5 interpretieren wir als Beleg für diese Klassenzahl, was zu einer Partitionierung von 93:69:50:139:119 führt.

Für *k-means* wird laut Abb. 3.8 jedoch die Klassenzahl 4 präferiert, weil sich erst kurz vorher die Interklassen- und die Intraklassen-Varianzsumme kreuzen und danach die Graphen knickfrei verlaufen. Die entsprechende Partitionierung ergibt sich zu 156:50:188:76 Sprechern.

Analyse von *gesMerkmVII*

Zur Veranschaulichung der hochdimensionalen Datenstruktur von *gesMerkmVII*, der ja aus neun verschiedenen Merkmalen besteht, ist eine Darstellung der beiden ersten Hauptkomponenten angezeigt.

Um die Frage der Klassenzahl für WARD *linking* zu entscheiden, betrachten wir Abb. 3.9; es stellt sich heraus, daß eine ähnliche Situation vorliegt wie bei *WortSatzIV*: Man sollte die im Diagramm hervorgehobenen Cluster 4 und 5 nicht zusammenfassen, weil die dabei

[14] Zumindest ist, wie bei *gesMerkmVa* , klar, daß die Klassenzahl größer als 3 sein muß.

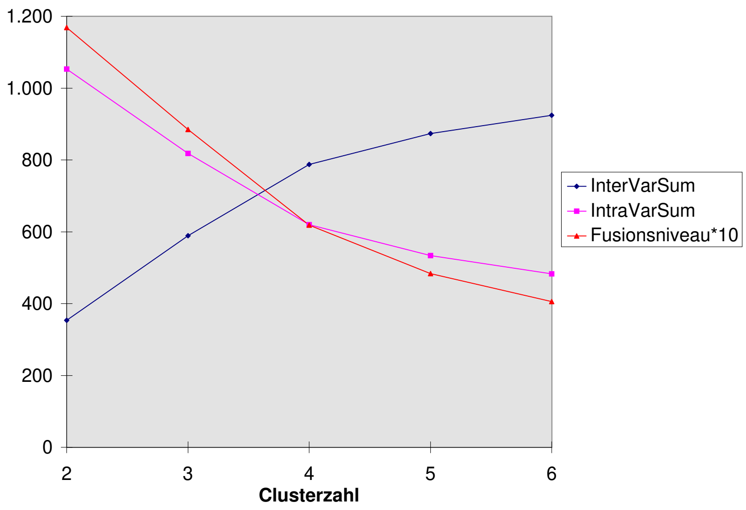

Abbildung 3.7: Varianzsummen und Fusionsniveaus von *gesMerkmVb*

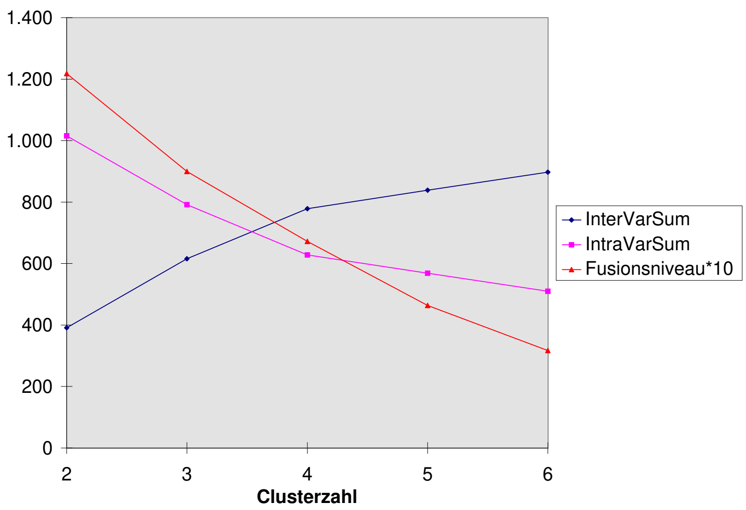

Abbildung 3.8: Varianzsummen und Fusionsniveaus von *gesMerkmVc*

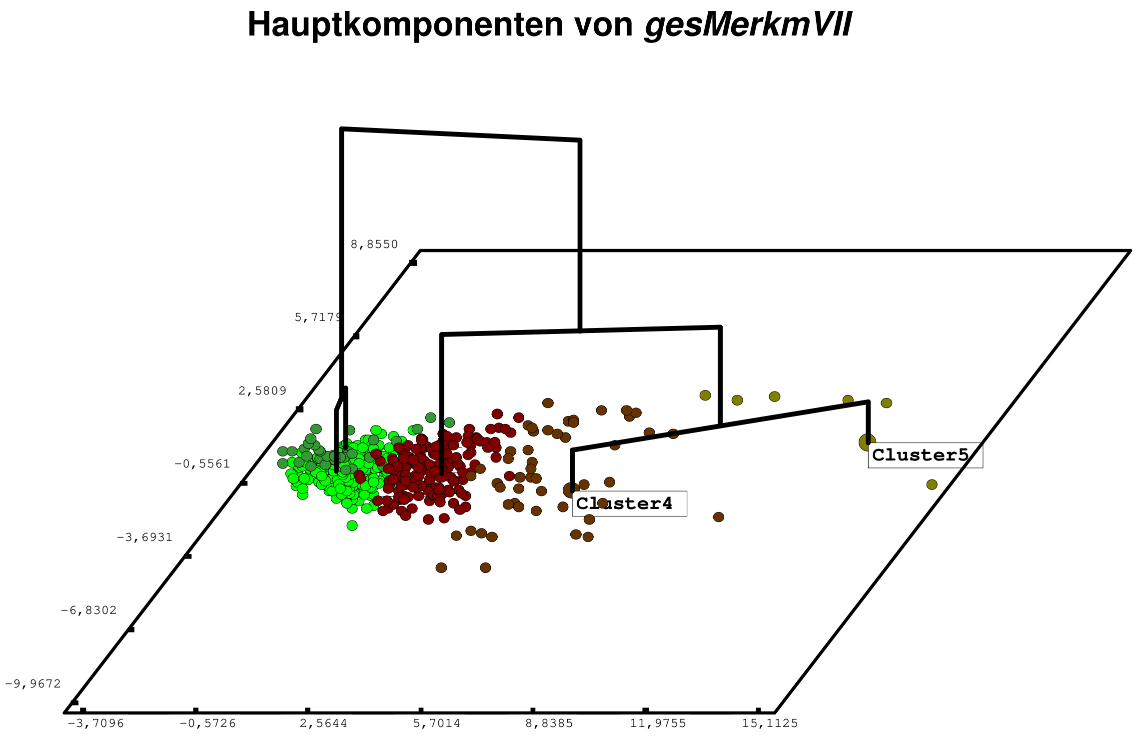

Abbildung 3.9: Hauptkomponenten-Dendrogramm von *gesMerkmVII*

entstehende Klasse zu heterogen wäre. Es ist also von 5 Klassen auszugehen. obwohl die „Klasse“ 5 diesen Namen im Grunde nicht verdient und nur ein Auffangbecken für Ausreißer unter den Daten darstellt[15].

Zu den Ergebnissen des Verfahrens *k-means* betrachten wir wieder eine grafische Darstellung der Varianzsummen (Abb. 3.10) und stellen dabei fest, daß keine bestimmte Klassenzahl > 3 favorisiert wird. Wir wählen deshalb für *k-means* dieselbe Zahl wie für Ward *linking* und erhalten die Partitionen 67:167:19:164:53 im ersten sowie 222:36:157:59:6 (man beachte die Größe der fünften „Klasse“) im zweiten Fall.

Analyse von *relEnde*

Der Merkmalssatz *relEnde* ist in mancher Hinsicht ein Sonderfall. Zum einen beruht seine Definition und nachfolgende Untersuchung auf einer visuellen Vorab-Klassifikation. Die Fra-

[15] Die Alternative wäre eine Entfernung der entsprechenden Sprecher aus der Stichprobe, was aber wegen der Vergleichbarkeit mit den anderen Partitionen für die Bewertung in Abschnitt 4.3 ausscheidet.

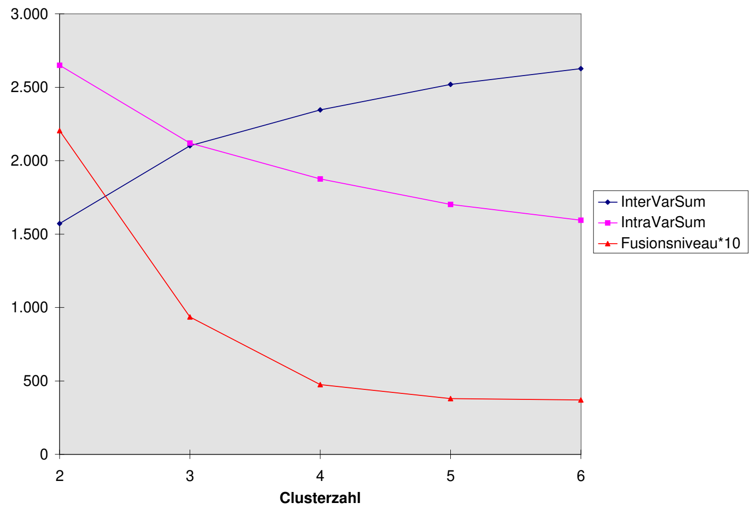

Abbildung 3.10: Varianzsummen und Fusionsniveaus von *gesMerkmVII*

ge nach der richtigen Klassenzahl stellt sich deshalb nicht, weil von vornherein drei Cluster vorgegeben waren, um die Ergebnisse der mathematischen Verfahren mit der visuellen Klassifikation zu vergleichen.

Zum andern enthält *relEnde* nur zwei Merkmale, und zwar relative Häufigkeiten; trotzdem wurde zur Klassifikation nicht, wie oben als Standardverfahren beschrieben, der χ^2-Abstand verwendet, sondern der EUKLIDische Abstand — wegen der geometrischen Interpretation, die ja den Ausgangspunkt für die Aufstellung von *relEnde* bildet.

Zum dritten schließlich wurde der Klassifikator[16] an diese geometrische Interpretation angepaßt und statt des sonst benutzten $3NN$-Klassifikators ein gewöhnlicher NN-Klassifikator verwendet.

Führt man die partitionierende Clusteranalyse *k-means* mit *relEnde* durch, so ergibt sich für drei Klassen jedoch ein ganz anderes als das aufgrund der visuellen Klassifikation erwartete Bild (Abb. 3.11). Hier zeigt sich wieder die Tendenz von *k-means*, mit dem Varianzkriterium kugelförmige Cluster zu bilden (vgl. [6], S. 184), die im $\Re^2$ natürlich als Kreise auftreten. Die Partitionierung von WARD *linking* jedoch ähnelt der visuellen Klassifikation (sie ist deshalb nicht explizit in die Darstellung eingefügt), wobei lediglich die Diagonale stärker ausgebildet ist.

Wir erhalten bei den für *relEnde* angewendeten drei Verfahren die Partitionierungen 139:177:154 für WARD *linking*, 194:145:131 für *k-means* und 47:285:138 für die visuelle Klassifikation.

Analyse von *Top10Bigr*

Die relativen Häufigkeiten der 10 häufigsten Bigramme wurden, wie angedeutet, ohne Standardisierung, aber mit dem Gewicht der inversen Standardabweichung und dem χ^2–Abstand

[16] diese Besonderheit betrifft die spätere Überprüfung und Bewertung der Klassifikation mit Hlfe von Teststichproben

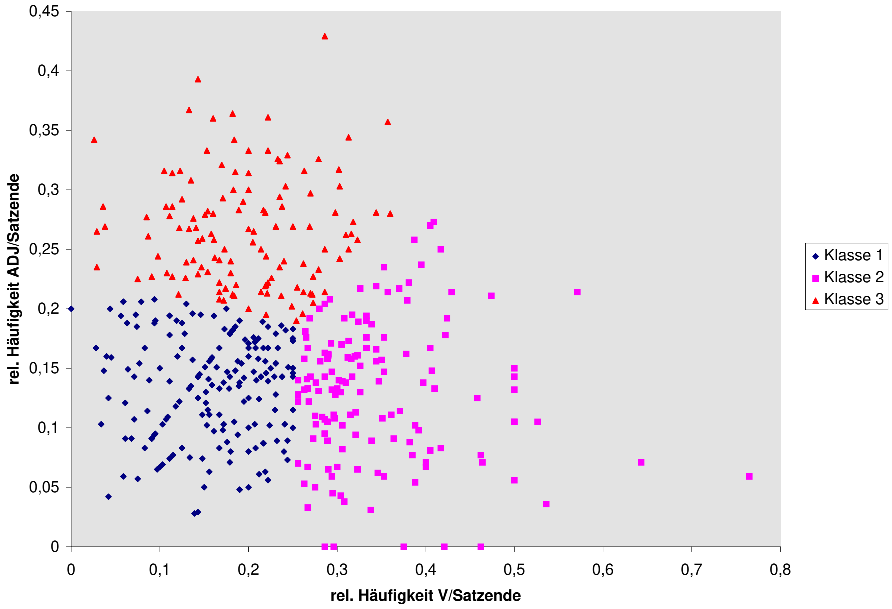

Abbildung 3.11: Clusterplot von *relEnde* mit *k-means*

als Distanzmaß der Clusteranalyse unterzogen.

Mit WARD *linking* ist jedoch keine klare Klassenstruktur auszumachen, wie das Hauptkomponenten-Dendrogramm in Abb. 3.13 zeigt; auch über die Klassenzahl kann daher keine eindeutige Aussage getroffen werden. Für *k-means* läßt sich in Abb. 3.12 aus dem Graphen des χ^2–Werts eine leichte Präferenz für die Klassenzahl 4 herauslesen, weshalb diese Klassenzahl auch für WARD *linking* übernommen wurde.

Die Partitionen ergeben sich damit zu 104:36:143:187 für die hierarchische und 135:110:122:103 für die partitionierende Clusteranalyse.

Schon jetzt ist jedoch absehbar, daß für *Top10Bigr* mit einem auch nur einigermaßen objektivierbaren Gütemaß kein gutes Resultat erreichbar sein wird; bei den anderen Partitionen wird das, so ist für die Bewertungen im nächsten Kapitel zu hoffen, nicht in gleichen Maße der Fall sein.

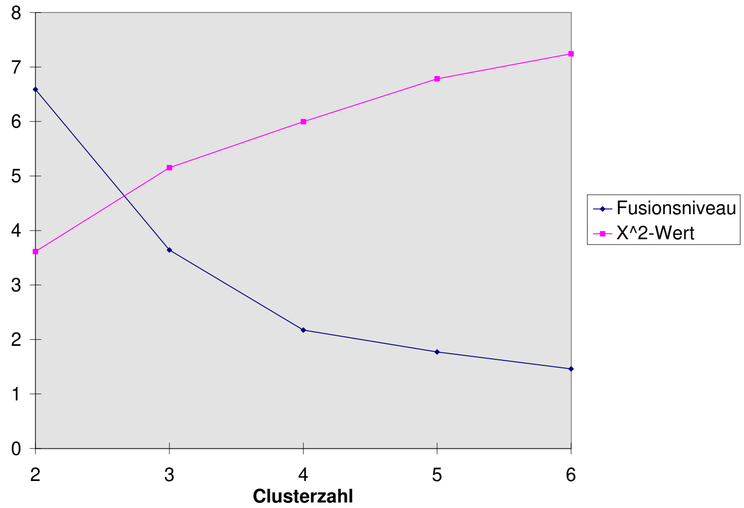

Abbildung 3.12: χ^2–Wert bei *k-means* und Fusionsniveau bei WARD *linking* von *Top10Bigr*

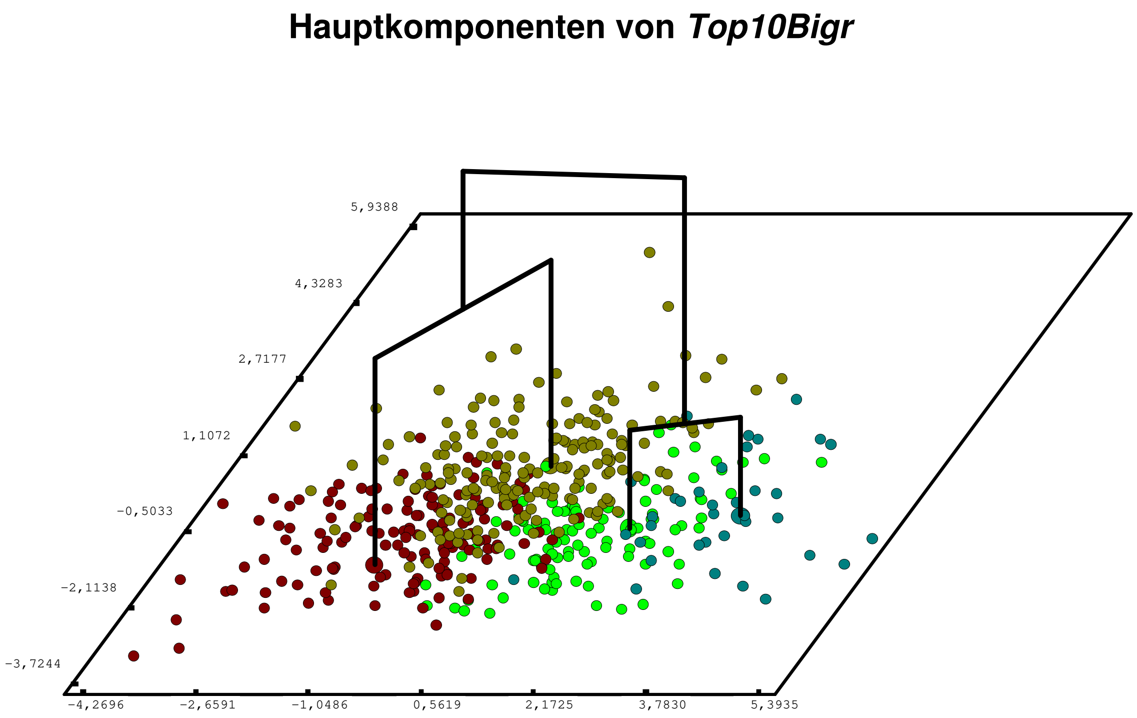

Abbildung 3.13: Hauptkomponenten-Dendrogramm von *Top10Bigr* mit χ^2-Abstand

Kapitel 4

Bewertung der Klassenstrukturen

Nach der Durchführung der Clusteranalysen, die in den letzten Abschnitten beschrieben wurde, sollen nun die ermittelten Klassen mit einem Sprachverarbeitungs- und damit anwendungsspezifischen Gütemaß, der Perplexität, bewertet werden. Dafür müssen wir jedoch zunächst einige theoretische Grundlagen ansprechen.

4.1 Sprachmodelle

Die Motivation für die Verwendung grammatischer Sprachmodelle in der Sprachverarbeitung liegt darin, daß die akustische Modellierung allein bei weitem zu viele Möglichkeiten in der Wortfolge offen läßt, wie ja auch ein menschlicher Sprecher stets auf verschiedene nicht-akustische Wissensquellen zurückgreift (vgl. [41], S. 447).

Beim statistischen Ansatz der linguistischen Sprachmodellierung ordnen stochastische Sprachmodelle jeder Wortfolge eine Wahrscheinlichkeit zu, die dann zusammen mit dem akustische Sprachmodell die Berechnung der a-posteriori-Wahrscheinlichkeit zuläßt; deren Maximierung ermöglicht den Rückschluß vom Sprachsignal auf die tatsächlich geäußerten Worte (vgl. die „Fundamentalformel der Spracherkennung“ in [35], S. 165). Die Wahrscheinlichkeit einer Wortfolge wird dabei unter der vereinfachenden, aber unumgänglichen Annahme berechnet, daß sie nur von einem begrenzten Kontext abhängt. Er umfaßt üblicherweise lediglich die letzten $1-2$ Wörter, was zu Bigramm- bzw. Trigramm-Modellen führt.

Die Wahrscheinlichkeit $P(\boldsymbol{w})$ einer Wortfolge $\boldsymbol{w} = w_1 \dots w_n$ ergibt sich bei Bigrammen bzw. Trigrammen zu

$$P(\boldsymbol{w}) = P(w_1) \cdot \prod_{i=2}^{m} P(w_i \mid w_{i-1}) \quad \text{bzw.} \quad P(\boldsymbol{w}) = P(w_1) \cdot P(w_2 \mid w_1) \cdot \prod_{i=3}^{m} P(w_i \mid w_{i-2} w_{i-1})$$

Der mittlere Verzweigungsfaktor innerhalb eines Wortmodells, also die Anzahl von dessen möglichen Fortsetzungen, wird als *Perplexität* bezeichnet. Mit den in [35], S. 206 angegebenen Abschätzungen ist die Perplexität $\hat{\wp}_x$ einer Trainingsfolge $\boldsymbol{w} = w_1 \dots w_m$ folgendermaßen definiert:

$$\hat{\wp}_x = 2^{\hat{\mathrm{H}}} = \hat{P}(\boldsymbol{w})^{-1/m}$$

$\hat{\mathrm{H}}$ ist dabei eine Abschätzung für die Entropie der Trainingsfolge, $\hat{P}(\boldsymbol{w})$ eine solche für die Produktionswahrscheinlichkeit der Wortfolge. Da diese maximiert werden soll (s. o.), sind demnach die Entropie bzw. die Perplexität zu minimieren.

Prinzipell haben Trigramm-Modelle auf ihrer Trainingsstichprobe geringere Perplexitäten als Bigramm-Modelle, weil durch den längeren Kontext die Anzahl möglicher Folgewörter geringer wird. Die Zahl der zu schätzenden Parameter des Wortmodells ist jedoch bei Trigrammen erheblich höher (bei n Vokabeln sind es n^2 Bigrammparameter und n^3 Trigrammparameter), so daß bei einer Teststichprobe aufgrund der wesentlich größeren Zahl ungesehener Parameter, d. h. unbekannter Wortfolgen, die Perplexität von Modellen mit längerem Kontext sogar schlechter sein kann als diejenige von Modellen mit kürzerem Kontext.

Und hier endlich kommt die Partitionierung der Trainingsmenge in Klassen, die mit Hilfe stilistischer Merkmale in den letzten Abschnitten ermittelt wurden, zum Tragen: Die in ihnen enthaltenen Sprecherdaten dienen als Trainingsstichproben für proprietäre Sprachmodelle. Diese Klassenmodelle weisen eine spezifische Verteilung für bestimmte Wortfolgen auf, die den speziellen Stil dieser Klasse ausmachen[1]. Man hofft nun, daß ein klassenspezifisches Sprachmodell einer Äußerung eines dem Modell unbekannten Sprechers aus der Teststichprobe, der dieser Klasse zugeordnet wird, eine höhere Wahrscheinlichkeit (und damit niedrigere Perplexität) zuweisen wird als das Gesamtmodell.

Der Effekt der ungesehenen Parameter tritt aber bei den klassenspezifischen Modellen gegenüber dem Gesamtmodell in qualitativ ähnlicher Weise auf wie bei Trigramm- gegenüber Bigramm-Modellen, weil ja dem Klassenmodell nur ein Teil der gesamten Trainingsdaten zur Verfügung stand (eben die Daten derjenigen Sprecher der Trainingsmenge, die diese Klasse bilden). Daraus folgt, daß die ungesehenen Wortfolgen den Gewinn an Perplexitätsreduktion, der durch die Spezialisierung erzielt wurde, teilweise oder sogar völlig zunichte machen können. Da ein Bigramm wesentlich weniger Parameter hat als ein Trigramm, wirkt sich der Effekt hier nicht so stark aus. Das bedeutet, daß man für Bigramme mit klassenspezifischen Sprachmodellen höhere Reduktionen der Perplexität erwarten kann als für Trigramme.

Die Zahl der zu schätzenden Parameter eines Wortmodells, also der in seinem Kontext möglichen Wortfolgen, wächst mit steigender Kontextlänge exponentiell an. Deswegen ist eine Reduktion der Parameter in den meisten Fällen ratsam[2]; sie kann durch die Einführung von Wortklassen erreicht werden.

Diese werden entweder automatisch ermittelt, beispielsweise auch mit den Mitteln der Clusteranalyse[3] oder nach linguistischen Kriterien vorgegeben. Im letzteren Fall handelt sich in der Regel um grammatische Kategorien, wie sie auch die POS-Tags im VM-Korpus darstellen. Die Kategorien müssen den Wörtern durchaus nicht immer eindeutig zugeordnet sein: Ein Wort kann, je nach Kontext, grundsätzlich mehrere syntaktische Funktionen erfüllen. Durch die Zuordnung der Kategorien im VM-Korpus mit dem POS-Tagger, die ja stets eindeutig erfolgt, kann man jedoch von disjunkten Kategorien ausgehen.

Bei disjunkten Kategorienbigrammen sind für m Kategorien bei einer Wortzahl n statt n^2 nur ungefähr $m^2 n$, bei Trigrammen $m^3 n$ Parameter zu schätzen; die fragliche Wahrscheinlichkeit $P(\boldsymbol{w})$ einer Wortfolge errechnet sich im Falle eines Modells mit disjunkten Kategorien dann zu

[1] Hier haben wir es mit einer extremen Eingrenzung des Stilbegriffs zu tun, die jedoch für den praktischen Einsatz in Sprachverarbeitungsystemen unumgänglich ist

[2] Die Verfahren zur Glättung von Verteilungen, also die Reservierung von Wahrscheinlichkeitsmasse für nicht in der Trainingsmenge enthaltene Wortfolgen, von denen es bei der Vielzahl der Parameter immer eine große Menge geben wird, sollen an dieser Stelle nicht weiter ausgeführt werden; auf Glättungsverfahren wird im Zusammenhang mit Ansätzen zur praktischen Verwendung von Stilklassen in Sprachverarbeitungssystemen noch kurz in 6.1 eingegangen.

[3] vgl.[42] für einen entsprechenden Ansatz zur Verallgemeinerung der Wortklassen auf Klassen von Wortfolgen

$$P(\boldsymbol{w}) = P(w_1) \cdot \prod_{i=2}^{m} P(w_i \mid C(w_i)) \cdot P(C(w_i) \mid C(w_{i-1}))$$

Wenn man die Perplexität eines Kategorienmodells berechnen will, muß man also eigentlich auch die Wahrscheinlichkeitsverteilungen der Wörter in den Klassen berücksichtigen. Man kann natürlich von der Hypothese ausgehen, daß die Wahrscheinlichkeiten $P(w_i \mid C(w_i))$ konstant sind und damit für die Berechnung irrelevant werden, aber das würde bedeuten, daß die Wörter in den kategorialen Wortklassen jeweils gleich verteilt sind — eine in der Praxis nicht zu rechtfertigende und nur als Näherung zulässige Annahme.

In dieser Arbeit wurden aus den genannten Gründen zur Verringerung des Effekts der ungesehenen Parameter Bigramm- und Trigramm-Kategorienmodelle verwendet, wobei die Perplexitäten jedoch mit dieser Näherung berechnet wurden[4].

4.2 Die Perplexität als Gütemaß

Obwohl bereits im vorigen Abschnitt die Perplexität als Gütemaß für diese Arbeit eingeführt wurde, sollen doch kurz einige andere Möglichkeiten für eine Bewertung der ermittelten Clusterpartitionen angesprochen werden.

Für jedes Verfahren der Clusteranalyse existieren anwendungsunabhängige, aber eben verfahrensspezifische Gütemaße, die als Kriterien durch das Verfahren optimiert werden. Für einen Vergleich der Resultate verschiedener Verfahren und Distanzmaße sind sie also nicht zu verwenden.

Andererseits lassen sich Klassenstabilität, Klassenzahl und die Klassenzuordnung allgemein in Simulationsuntersuchungen überprüfen, bei denen zufällig generierte Stichproben einer Clusteranalyse mit dem zu überprüfenden Verfahren unterzogen werden (vgl. [28], S. 151 ff.). Aus einem Vergleich von dabei ermittelten Referenzwerten mit den entsprechenden Werten aus der Clusteranalyse der tatsächlich vorhandenen Daten lassen sich Rückschlüsse auf die Güte und Relevanz der Klassifikation ziehen. Dieses Verfahren ist jedoch aufwendig und für den konkreten Nutzeffekt der Ergebnisse in der Sprachverarbeitung nicht aussagekräftig.

Im Zweifelsfall ist natürlich einem Merkmalssatz und einer Partition mit geeigneter linguistischer Interpretation der Vorzug zu geben; überdies wird eine noch so gute Klassifikation der Sprecher, wenn sie auf einem in linguistischer Hinsicht völlig interpretationslosen Merkmalssatz[5] beruht, immer ein unbefriedigendes Ergebnis darstellen. Man hat hier eine zwar anwendungsspezifische, aber doch nur qualitative Bewertungsmöglichkeit vorliegen. Für den Praktiker, d. h. den Entwickler und vor allem den Anwender von Sprachverarbeitungssystemen, werden Überlegungen zur linguistischen Relevanz allerdings eine untergeordnete Rolle spielen.

Denn jenseits aller Kriterien zur Merkmalsauswahl und Klassifikationsbewertung, aller anwendungsunabhängigen lokalen und globalen Gütemaße und jeglicher theoretisch interessanten linguistischen Interpretation steht die Bewährung in der Praxis, m. a. W. der Nutzen

[4] Der praktische Grund dafür ist darin zu sehen, daß die Perplexitätsberechnungen mit den *CMU-Tools* vorgenommen wurden, die auf den in [23] und [22] dargestellten Verfahren und Resultaten von KATZ und JELINEK basieren: Sie sind eigentlich für Wortmodelle vorgesehen und geben daher die angesprochene Näherung vor.

[5] man denke an BOLZ' Negativbeispiel „Anzahl der Wörter mit dem Buchstaben t an zweiter Stelle“, vgl. [7], S. 205

einer Untersuchung und Klassifikation persönlicher Sprachstile zeigt sich erst im Einsatz eines Sprachverarbeitungssystems. Während die in [35], S. 341 ausgesprochene Vermutung hinsichtlich der datengetriebenen Verbesserungen von Spracherkennern gewissermaßen als Motivation für die Untersuchungen in dein vorangegangenen Abschnitten gelten kann, findet sich beim selben Autor auch die Begründung, warum im folgenden die Perplexität als das „Maß aller Dinge“ für den praktischen Wert dieser Untersuchungen herangezogen wird: Nach [35], S. 340 zeigt sich „eine signifikante Korrelation zwischen der erzielten Erkennungssicherheit und der Stichprobenperplexität“ bei verschiedenen spontansprachlichen Stichproben, und einer "Daumenregel zufolge [sinkt] die Wortfehlerrate [...] etwa um die Quadratwurzel des Reduktionsfaktors“ ([35], S. 206).

4.3 Bewertungsverfahren

4.3.1 Trainings- und Teststichprobe

Der Nutzen der Perplexität zur Bewertung der Merkmalssätze und Partitionen wurde in den vorangegegangen Abschnitten dargestellt; wie sieht nun das Vorgehen im einzelnen aus?

Als erster Schritt wird die Trainingsmenge durch die jeweils zu bewertende Klassifikation in Untermengen partitioniert, die durch die Stilklassen gebildet werden. Für jede ermittelte Klasse wird ein Sprachmodell mit den Daten der Klassenmitglieder erstellt, außerdem ein Sprachmodell der gesamten Trainingsstichprobe. Laut der dieser Arbeit zugrundeliegenden Arbeitshypothese III sollten bei einem bisher unbekanntem Sprecher, dessen Daten in keinem gegebenen Modell verarbeitet wurden, mit dem seiner Stilklasse zugeordneten Sprachmodell bessere Erkennungsraten erzielt werden als mit den Sprachmodellen der anderen Klassen und insbesondere mit demjenigen der gesamten Stichprobe.

Zur Überprüfung dieser Hypothese mit der Kenngröße Perplexität muß zunächst eine Teststichprobe festgelegt werden. Dazu wurden CD 2 und als weitere Teststichprobe CD 4 getrennt voneinander herangezogen.

Um die Sprecher der Teststichprobe den Stilklassen der Trainingsmenge zuzuordnen, benötigt man einen Klassifikator. Hier eignet sich *kNN*-Klassifikation, wobei zweckmäßig dasselbe Distanzmaß zur Anwendung kommt, das bereits bei der Clusteranalyse verwendet wurde. Bei *kNN*-Klassifikation mit ungeradem k läßt sich im Falle mehrerer gleicher Abstände zu Punkten aus verschiedenen Klassen stets eine Majoritätsentscheidung herbeiführen; in dieser Arbeit wurde ein *3NN*-Klassifikator verwendet[6].

4.3.2 Bewertung über Anteile optimal klassifizierter Sprecher

Es bieten sich nun zwei unterschiedliche Vorgehensweisen bei der Bewertung der Stilklassen mit der Perplexität an.

Der erste Weg, ein relativer Vergleich der verschiedenen Merkmalssätze und Partitionen untereinander, führt über eine Rangliste bezüglich des prozentualen Anteils an optimalen Zuordnungen, d. h. solchen, bei denen ein Sprecher der Klasse zugewiesen wird, mit deren Sprachmodell die Perplexität seiner Daten den minimalen Wert unter den verfügbaren Perplexitäten der verschiedenen Klassen annimmt.

[6] Für den hypothetischen Fall exakt gleicher Abstände zu drei Sprechern aus drei verschiedenen Klassen wurde eine zufällige Zuordnung vorgenommen.

Man berechnet dazu die Perplexitäten jedes Sprechers der Teststichprobe mit allen verfügbaren Sprachmodellen, also dem der gesamten Trainingsstichprobe und denjenigen der einzelnen Klassen. Hat man einen Sprecher hinsichtlich seiner Perplexität optimal klassifiziert, so sollte die Perplexität in diesem Fall den minimalen aller berechneten Sprecherwerte annehmen.

Das Verfahren wird für alle zu bewertenden Merkmalssätze und Partitionen, die durch die Verfahren *k-means* und WARD *linking* errechnet wurden, durchgegeführt. Mit den dabei gewonnenen Daten über die Anzahl optimal klassifizierter Sprecher läßt sich eine Rangliste der Merkmalssätze und Verfahren aufstellen.

Ein Nachteil des Verfahrens ist allerdings darin zu sehen, daß es geringere Klassenzahlen bevorzugt: Generell sinkt die Chance, daß die Perplexität bei der Sprecherklasse ein Minimum annimmt, mit der Zahl der Klassen, und so werden Partitionen mit z. B. zwei Klassen stets besser bei einer Zählung der in dieser Hinsicht korrekt klassifizierten Sprecher abschneiden als Partitionen mit vier Klassen. Eine Bewertung der Partitionen kann also nur nach der Klassenzahl getrennt erfolgen. Andererseits hat es den Vorteil, nicht vom Sprachmodell abzuhängen, so daß Bigramm- und Trigramm-Modelle gemeinsam bewertet werden können.

Am günstigsten ist dabei wohl der Fall anzusehen, bei dem die Perplexitäten der anderen Klassen noch über der Perplexität der gesamten Trainingsstichprobe liegen: Die Trennschärfe der Klassenzuordnung wäre damit eindrucksvoll belegt. Dabei ist jedoch das o. a. Parameterproblem zu berücksichtigen: Die Klasseneinteilungen sind also nicht allein schon deshalb abzulehnen, weil mit dem globalen Sprachmodell in manchen Fällen geringere Perplexitätswerte erreicht werden als mit den klassenspezifischen Modellen.

Dies führte zur Idee, „Chancengleichheit“ bei den Sprachmodellen durch jeweils gleich große zur Verfügung stehende Datenmengen herzustellen[7]. Dazu wurden aus der Trainingsmege Stichproben in Klassengröße zufällig ausgewählt und mit diesen Stichproben ebenfalls Sprachmodelle trainiert[8].

Ein weiterer Schritt ist die Umgehung des NN-Klassifikators oder vielmehr seine Ersetzung durch die „Klassifikationsvorschrift“, einen unbekannten Sprecher der Klasse zuzuordnen, mit deren Sprachmodell sein Perplexitätswert minimal wird, wodurch dieser von einer Kontrollgröße für die Klassifikation zur Klassifikationsgröße selbst wird. Der Vorteil liegt auf der Hand: Dem Sprecher wird stets das beste im Rahmen der Stilklassen verfügbare Sprachmodell zugewiesen. Der so erhaltene Wert kann demnach als theoretisches Optimum der jeweiligen Klassifikation angesehen werden.

4.3.3 Bewertung über die Stichprobenperplexität

Die im letzten Abschnitt beschriebene Bewertung der Klassifikation mit einer relativen Rangordnung unter den Merkmalssätzen, Partitionen und Verfahren aufgrund des Anteils der Zuordnungen zur Klasse mit der minimalen Perplexität ist aber nur dann von Interesse, wenn auch eine Reduktion der Stichprobenperplexität gelingt; denn was nützt eine relativ beste Klassifikation, wenn die Perplexität des Gesamtmodells noch deutlich unter derjenigen der besten Klasse liegt?

[7]„Gleich groß“ ist in diesem Fall ein dehnbarer Begriff, der von „Anzahl der Wörter“ bis „Anzahl der Sprecher“ reichen kann; als ein vernünftiger Kompromiß zwischen exakter quantitativer Übereinstimmung und grammatikalisch (d.h. für ein aussagefähiges Sprachmodell) sinnvoller Aufspaltung wurde die Anzahl der Turns gewählt.

[8]Das Verfahren zeigt eine strukturelle Ähnlichkeit mit den Simulationsstudien auf Zufallsdaten zur Bewertung von Klassifikationen.

Zur Bewertung mit der Stichprobenperplexität bestimmt man zuerst die Perplexität der aus den Daten der Sprecher in der Teststichprobe, die der jeweiligen Klasse zugeordnet sind, zusammengesetzten Wortfolge, mit dem Sprachmodell der entsprechenden Klasse der Trainingsstichprobe. Die Stichprobenperplexität $\wp_x^s$ wird dann folgendermaßen aus den logarithmierten Produktions-Wahrscheinlichkeiten der k einzelnen Stilklassenmodelle errechnet[9]:

$$\mathrm{H} = -\frac{1}{m} \cdot \sum_{i=1}^{k} (-m_i \log_2 \wp_{xi}) \quad \text{und} \quad \wp_x^s = 2^{\mathrm{H}}$$

m ist dabei die Gesamtzahl der Wörter in der Teststichprobe, m_i die Wortzahl der (Teststichproben-)Klasse i; die Summation über den Logarithmus erfolgt aus technischen Gründen, weil die Produkte über die Wahrscheinlichkeiten schnell sehr klein werden.

Dann wird die Stichprobenperplexität mit der Perplexität der gesamten Trainingsmenge verglichen und überprüft, ob und inwieweit eine Reduktion beim speziellen Modell gegenüber dem globalen erreicht wurde.

4.4 Ergebnisse der Merkmalssätze und Partitionen

Bei der Umsetzung des skizzierten Verfahrens in die Praxis zeigte sich allerdings eine unangenehme Eigenheit der *CMU-Tools*, die zur Berechnung der einzelnen Perplexitäten mit den klassenspezifischen Sprachmodellen und dem Gesamtmodell verwendet wurden, ohne ersichtlichen Grund in manchen Fällen bei der Perplexitätsberechnung mit einer Fehlermeldung abzubrechen.

4.4.1 Resultate mit prozentualen Anteilen

Für das erste skizzierte Verfahren der sprecherbezogenen Bewertung über den Anteil an Zuordnungen zur Klasse mit der minimalen Perplexität ergab sich daher in einigen Fällen die Situation, daß gerade für die Klasse, der ein Sprecher durch den Klassifikator zugeordnet werden sollte, kein Wert zur Verfügung stand. In diesen Fällen wurde die fehlende Perplexität des Sprechers durch das Maximum der Werte mit den Modellen der anderen Klassen ersetzt. Dieses Vorgehen entspricht also einer „Strafversetzung“ des Sprechers in die schlechteste Klasse.

Gehen wir zunächst noch kurz auf den Vergleich der Perplexitätswerte von Modellen aus zufälligen Stichproben in Klassengröße mit den tatsächlichen Klassenmodellen ein.

Während beim direkten Vergleich der Sprachmodelle tatsächlich in der weitaus überwiegenden Zahl der Fälle das globale Modell am besten abschneidet, ergibt sich bei dieser Methode, ein vom Umfang „faires“ Sprachmodell einer Zufallsstichprobe gleicher Länge wie die jeweilige Klasse zu verwenden, ein ganz anderes Bild: Die Werte der Sprecherklasse liegen fast durchgehend unter denjenigen der Stichproben. Lediglich in den Fällen fehlerhafter Klassifikation im Sinne der Perplexität, in denen der Sprecher einer Klasse zugewiesen wurde, die im Vergleich zu den Sprachmodellen der anderen Klassen einen hohen, wenn nicht sogar den maximalen Perplexitätswert aufweist, liegt der Wert bei dieser offensichtlich suboptimalen Klasse über derjenigen der Stichprobe.

Aus diesen Ergebnissen läßt sich aber nur der begrenzt aussagefähige Schluß ziehen, daß die Partitionen zumindest bessere Resultate liefern als eine zufällige Auswahl; wenden wir

[9] Wir nehmen im folgenden stillschweigend den Schätzwertcharakter der ermittelten Wahrscheinlichkeit, Entropien und Perplexität an und ersparen uns die entsprechende Kennzeichnung der Werte wie in $\hat{\wp}_x$.

Tabelle 4.1: Zuordnung in die Klasse minimaler Perplexität mit CD 2

Merkmalssatz	Verfahren	Modell	Clusterzahl	Anzahl	Gesamtzahl	Prozent
relEnde	*k-means*	Trigramm	3	59	77	**76,62**
WortSatzIV	WARD	Bigramm	4	50	79	**63,29**
gesMerkmVII	*k-means*	Trigramm	5	46	74	**62,16**
WortSatzIIa	*k-means*	Trigramm	3	44	77	57,14
WortSatzIV	WARD	Trigramm	4	44	77	57,14
gesMerkmVII	*k-means*	Bigramm	5	44	76	57,89
gesMerkmVII	WARD	Trigramm	5	43	77	55,84
relEnde	*Assign*	Trigramm	3	34	77	44,16

Tabelle 4.2: Zuordnung in die Klasse minimaler Perplexität mit CD 4

Merkmalssatz	Verfahren	Modell	Clusterzahl	Anzahl	Gesamtzahl	Prozent
relEnde	*k-means*	Trigramm	3	55	73	**75,34**
WortSatzIV	WARD	Trigramm	4	47	73	**64,38**
WortSatzIIa	*k-means*	Trigramm	3	46	73	**63,01**
gesMerkmVII	WARD	Bigramm	5	45	74	**60,81**
gesMerkmVII	WARD	Trigramm	5	44	73	**60,27**
WortSatzVa	*k-means*	Trigramm	5	39	65	60,00
relEnde	*Assign*	Trigramm	3	41	73	56,16
WortSatzIV	WARD	Bigramm	4	40	74	54,05

uns daher jetzt der Bewertung der prozentualen Anteile optimal, d. h. mit minimaler aus den Klassen verfügbarer Perplexität klassifizierter Sprecher zu. In den Tabellen 4.1 und 4.2, die nach den Prozentwerten sortiert sind, wurden die Werte über 50% eingetragen; falls eine Zuordnung zur Klasse mit der minimalen Perplexität in weniger als 50% der Fälle erfolgt, ist die Kombination aus dem betreffenden Merkmalssatz und Verfahren irrelevant und braucht nicht weiter berücksichtigt werden.

Aus den Tabellen ist zu entnehmen, daß die Ergebnisse bei CD 4 etwas besser sind als bei CD 2. Von der Rangfolge der Merkmale ergeben sich keine großen Unterschiede zwischen den beiden Teststichproben: In den vorderen Rängen findet sich die Gruppe aus Wort-, Satz- und Turnlängen sowie *gesMerkmVII*. Hinsichtlich der Verfahren und auch der Modelle bestehen augenscheinlich keine Präferenzen.

Überraschend ist der Anteil von 75% bei *relEnde* mit dem Verfahren *k-means*, interessant der Wert von immerhin noch 56% bei der visuellen Klassifikation, die wir hier mit *Assign*[10] bezeichnet haben; dieses relativ gute Ergebnis wird aber, wie sich gleich herausstellen wird, von den Ergebnissen der Stichprobenperplexität nicht bestätigt werden können.

4.4.2 Resultate unter Verwendung der Stichprobenperplexität

Befassen wir also nun mit der Ermittlung der praktisch relevanten Stichprobenperplexitäten.

[10] in Anlehnung an ein ebenfalls in *ClusCorr* verfügbares partitionierendes Verfahren, das aber so schlechte Resultate bei der Klassifikation lieferte, daß es in der vorliegenden Arbeit abgesehen von dieser Stelle keine Erwähnung mehr findet

Dabei stellte sich jedoch heraus, daß die Werte nicht in aussagefähiger Weise direkt auf Clusterbasis zu berechnen waren: In manchen Fällen war nicht einmal mit der o. a. „Strafversetzung" ein clusterbezogener Wert zu erhalten, weil die CMU-Berechnung für alle Cluster der Teststichprobe abbrach[11].

Es bot sich deshalb der Weg über die einzelnen Sprecher an, deren Perplexitäten ja ohnehin für die erste Variante der Bewertung über den prozentualen Anteil der Zuordnung zur Klasse mit der minimalen Perplexität herangezogen wurde, weil der Ausfall einzelner Sprecher nicht so ins Gewicht fällt wie derjenige ganzer Cluster.

Man muß sich allerdings im klaren sein, was das für die Berechnung der Stichprobenperplexität bedeutet; sie ergibt sich jetzt nämlich zu:

$$\mathrm{H} = -\frac{1}{m} \cdot \sum_{i=1}^{s} (-m_i' \log_2 \wp_{xi}') \quad \text{und} \quad \wp_x^s = 2^{\mathrm{H}}$$

s ist dabei die Anzahl der Sprecher in der Teststichprobe, m_i' die Wortzahl des Sprechers i und $\wp_{xi}'$ die Perplexität des Sprechers i mit dem Sprachmodell seiner Klasse.

Die Teststichprobe wird also in s Sprecher- statt k Klassen-Partitionen zerlegt; natürlich existieren weiterhin nur die k Sprachmodelle aus der Trainingsmenge.

Im Falle von Bigrammen wirkt sich die Verfeinerung der Partitionen auf die Sprecherebene und die daraus resultierende Entfernung der n-Gramme für die Übergänge von einem Sprecher zum nächsten nicht weiter aus, weil lediglich die Anzahl des Bigramms „<Satzende> <Satzanfang>" um $n-k$ abnimmt. Als Konsequenz wird die Wahrscheinlichkeit, daß auf ein Satzende eine neuer Satzanfang folgt, unwesentlich geringer; sie wird aber immer noch sehr nahe bei 1 liegen. — Bei Trigrammen werden immerhin an den Übergängen zwischen den Sprechern alle Folgen „X <Satzende> <Satzanfang>" und „<Satzende> <Satzanfang> X" enfernt, also $O(m)$ Stück (m ist die Größe des Vokabulars). Doch ist ja bei Trigrammen auch die gesamte Parameterzahl wesentlich größer, so daß der Effekt nicht weiter ins Gewicht fällt.

Die gleiche Argumentation läßt sich für die Perplexität der Teststichprobe mit dem allgemeinen Modell anführen. Um eine Vergleichbarkeit bei der Wortzahl zu gewährleisten, muß dieser Perplexitätswert jeweils auf dem Teil der Stichprobe berechnet werden, der auch für die Stichprobenperplexität der Cluster zur Verfügung stand. Statt also einen Wert für die gesamte Stichprobe zu berechnen, wurde nach dem gleichen Schema zunächst die Perplexität jedes Sprechers mit dem Gesamtmodell berechnet und die Stichprobenperplexität dann wie oben bestimmt[12].

Werfen wir für genauere Aussagen zu den Ergebnissen der Stichprobenperplexität nun einen Blick auf die Übersichtstabellen 4.3, 4.5, 4.4 und 4.6; die Einträge sind nach aufsteigender Stichprobenperplexität $\wp_x^s$ sortiert[13] und enthalten neben der Perplexität des Gesamtmodells $\wp_x^{ges}$ auch den minimalen mit der jeweiligen Partition erreichbaren Wert $\wp_x^{min}$, zu dessen Berechnung jeweils die minimale Perplexität einer Sprechers unter allen Stilklassen herangezogen wurde.

Für den Leser, der mehr an qualitativen Aussagen über die Merkmalssätze und Verfahren mit den verschiedenen Sprachmodellen in der Anwendung auf die Teststichproben interessiert

[11]Dazu genügt es, wenn in jedem Cluster der Teststichprobe ein Sprecher enthalten ist, für den die Berechnung abbricht.

[12]ein Verfahren, dessen Beschreibung sich aufwendig liest, das aber wegen der Parallelität zur clusterspezifischen Berechnung der Stichprobenperplexität ganz einfach zu realisieren ist

[13]Wenn der Leser also den einen oder anderen Merkmalssatz oder Verfahren vermissen sollte, so deshalb, weil nur die Einträge mit den niedrigeren Perplexitätswerten aufgeführt sind — mit Ausnahme von *Top10Bigr*.

ist, sind die Ergebnisse in Abb. 4.1 gesammelt und veranschaulicht.

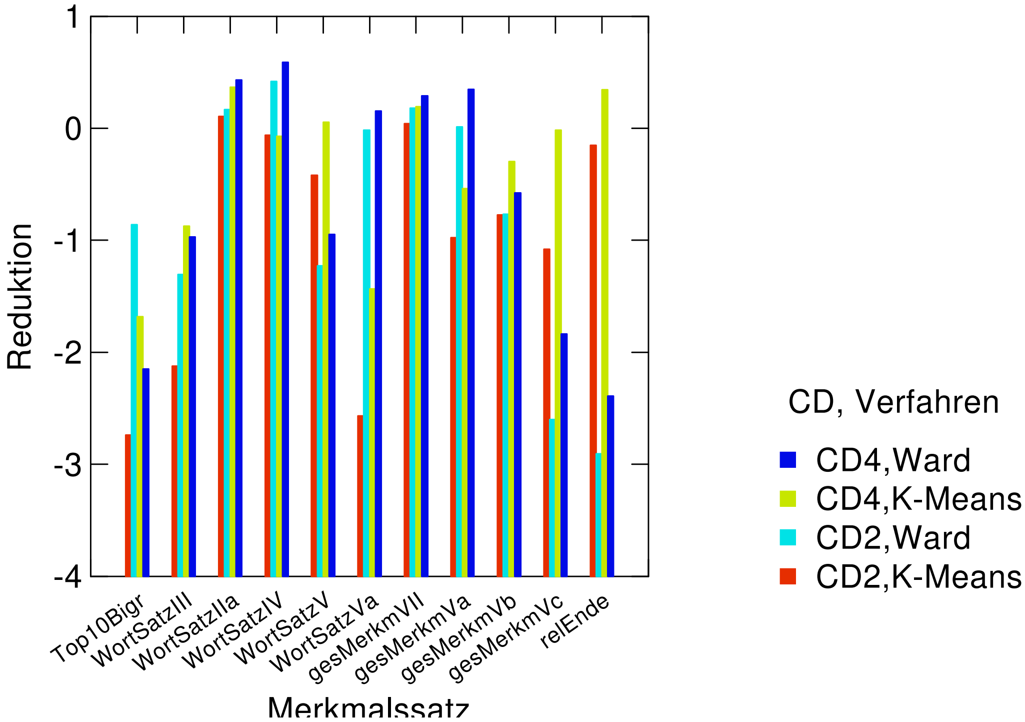

Abbildung 4.1: Reduktionen der Stichprobenperplexität im Vergleich

Man erkennt zunächst, daß die Erwartungen in 4.3 bezüglich der Reduktionen eingetreten sind, nicht jedoch im Hinblick auf die absoluten Werte: Die Bigrammodelle weisen die höheren absoluten Werte, aber auch die höheren Reduktionen der Stichprobenperplexität auf.

Außerdem ist zu sehen, daß CD 4 deutlich besser abschneidet als CD 2, was die Perplexitätsreduktion bei den besten Merkmalssätzen angeht: Zum einen ist $\wp_x^{ges}$ bei CD 2 etwas geringer, zum anderen ist $\wp_x^s$ bei CD 2 stets etwas höher, so daß die Differenz von beiden Seiten aus reduziert wird. Damit wird bei den Trigramm-Modellen mit CD 2 nur einmal eine Reduktion der Gesamt-Perplexität erreicht, nämlich für *WortSatzIV* mit WARD *linking*. Bei CD 4 hingegen ist auch *WortSatzIIa* etwas besser als $\wp_x^{ges}$, und sogar *relEnde* führt zu einer leicht geringeren $\wp_x^s$.

Betrachtet man die Ergebnisse in Bezug auf die Clusteranalyseverfahren, so ergibt sich fast stets ein Vorteil für WARD *linking* gegenüber *k-means*; lediglich die offenkundig schlechten Merkmalssätze erbringen mit WARD *linking* noch schlechtere Werte (s. z. B. *WortSatzIII*); die einzige Ausnahme stellt wieder einmal *relEnde* dar: In diesem Fall zeigt sich die Clusteranalyse mit *k-means* überlegen gegenüber der visuellen Klassifikation und auch gegenüber WARD *linking*, das eine dieser Klassifikation ähnliche, aber offenbar für die Anwendung doch nicht passende Klassenstruktur ergeben hat; wir werden dies bei der linguistischen Interpretation zu berücksichtigen haben.

Das hierarchische Verfahren WARD *linking* differenziert also Merkmalssätze und Partitio-

Tabelle 4.3: Stichprobenperplexität von CD 2 mit Bigramm-Kategoriemodellen

Merkmalssatz	Verfahren	Cluster	$\wp_x^s$	$\wp_x^{min}$	$\wp_x^{ges}$	Sprecher	Wörter
gesMerkmVa	WARD	4	10,45	10,33	10,61	47	12462
WortSatzIV	WARD	4	**10,90**	10,79	11,51	79	32680
gesMerkmVII	WARD	5	**11,08**	10,84	11,51	79	32680
WortSatzIIa	WARD	3	**11,12**	10,91	11,51	79	32680
WortSatzIIa	*k-means*	3	**11,15**	10,87	11,51	79	32680
gesMerkmVII	*k-means*	5	**11,24**	10,91	11,56	76	30686
WortSatzVa	WARD	4	**11,32**	10,82	11,51	79	32680
WortSatzIV	*k-means*	4	**11,36**	10,91	11,51	79	32680
relEnde	*k-means*	3	11,53	11,03	11,51	79	32680
Top10Bigr	WARD	4	12,31	10,95	11,51	79	32680
gesMerkmVa	*k-means*	4	12,41	10,91	11,51	79	32680
WortSatzIII	*k-means*	4	12,72	10,86	11,51	79	32680
WortSatzIII	WARD	4	12,85	10,95	11,51	79	32680
relEnde	Assignment	3	*13,61*	11,79	11,51	79	32680
WortSatzVa	*k-means*	4	13,81	11,02	11,51	79	32680
Top10Bigr	*k-means*	4	14,37	11,26	11,51	79	32680
relEnde	WARD	3	15,13	12,78	11,51	79	32680

Tabelle 4.4: Stichprobenperplexität von CD 4 mit Bigramm-Kategoriemodellen

Merkmalssatz	Verfahren	Cluster	$\wp_x^s$	$\wp_x^{min}$	$\wp_x^{ges}$	Sprecher	Wörter
gesMerkmVa	WARD	4	10,78	10,72	11,40	*39*	12603
WortSatzIV	WARD	4	**10,94**	10,82	11,77	74	31682
WortSatzIIa	WARD	3	**11,06**	10,86	11,77	74	31682
gesMerkmVII	WARD	5	**11,15**	10,83	11,77	74	31682
WortSatzIIa	*k-means*	3	**11,17**	10,86	11,77	74	31682
relEnde	*k-means*	3	**11,18**	10,96	11,77	74	31682
gesMerkmVII	*k-means*	5	**11,24**	10,96	11,77	74	31682
WortSatzVa	WARD	4	**11,41**	10,85	11,77	74	31682
gesMerkmVc	*k-means*	4	**11,44**	10,89	11,77	74	31682
WortSatzIV	*k-means*	4	**11,65**	10,92	11,77	74	31682
gesMerkmVb	*k-means*	4	11,88	10,90	11,77	74	31682
WortSatzIII	*k-means*	4	12,60	10,85	11,77	74	31682
WortSatzIII	WARD	4	12,78	11	11,77	74	31682
WortSatzVa	*k-means*	4	13,11	10,95	11,77	74	31682
relEnde	Assign	3	13,31	11,86	11,77	74	31682
Top10Bigr	*k-means*	4	13,37	11,17	11,77	74	31682

Tabelle 4.5: Stichprobenperplexität von CD 2 mit Trigramm-Kategoriemodellen

Merkmalssatz	Verfahren	Cluster	$\wp_x^s$	$\wp_x^{min}$	$\wp_x^{ges}$	Sprecher	Wörter
gesMerkmVa	WARD	4	8,82	8,68	8,69	*53*	*15065*
WortSatzIV	WARD	4	**9,46**	9,26	9,69	77	31953
WortSatzIIa	WARD	3	9,74	9,47	9,69	77	31953
gesMerkmVII	WARD	5	9,76	9,35	9,69	77	31953
WortSatzIIa	*k-means*	3	9,84	9,34	9,69	77	31953
WortSatzVa	WARD	4	9,91	9,36	9,69	77	31953
gesMerkmVII	*k-means*	5	9,92	9,62	9,68	74	29959
WortSatzIV	*k-means*	4	9,97	9,54	9,69	77	31953
relEnde	*k-means*	3	9,98	9,60	9,69	77	31953
WortSatzVa	*k-means*	5	10,31	9,40	9,69	77	31953
WortSatzVa	WARD	5	11,04	9,56	9,69	77	31953
relEnde	Assignment	3	11,48	10,04	9,69	77	31953
WortSatzVa	*k-means*	4	12,53	9,68	9,69	77	31953

Tabelle 4.6: Stichprobenperplexität von CD 4 mit Trigramm-Kategoriemodellen

Merkmalssatz	Verfahren	Cluster	$\wp_x^s$	$\wp_x^{min}$	$\wp_x^{ges}$	Sprecher	Wörter
gesMerkmVa	WARD	4	**9,46**	9,31	9,52	*48*	*15777*
WortSatzIV	WARD	4	**9,46**	9,25	9,82	73	30560
WortSatzIIa	WARD	3	**9,66**	9,37	9,82	73	30560
WortSatzIIa	*k-means*	3	**9,67**	9,31	9,82	73	30560
relEnde	*k-means*	3	**9,72**	9,47	9,82	73	30560
WortSatzVa	WARD	4	9,85	9,35	9,82	73	30560
gesMerkmVII	WARD	5	9,86	9,26	9,82	73	30560
gesMerkmVII	*k-means*	5	9,95	9,59	9,82	73	30560
WortSatzVa	*k-means*	5	10,07	9,30	9,84	65	26780
WortSatzIV	*k-means*	4	10,08	9,50	9,82	73	30560
gesMerkmVc	*k-means*	4	10,17	9,41	9,82	73	30560
WortSatzIII	*k-means*	4	10,73	9,40	9,82	73	30560
relEnde	Assignment	3	11,80	10,09	9,82	73	30560

Tabelle 4.7: Clusterverteilungen von CD 4 mit *WortSatzVa*

Cluster	*k-means*	WARD *linking*
1	**43**	24
2	**0**	15
3	6	10
4	6	10
5	10	14

nen stärker aus als das partitionierende *k-means*. Was könnte aber der Grund dafür sein, das mit *k-means* erstellte Partitionen bei relativ guten Merkmalssätzen schlechtere Perplexitätswerte liefern?

Es wurde schon mehrfach darauf hingewiesen, daß *k-means* zur Bildung kugelförmiger Cluster tendiert. Dies kann aber in den Fällen, in denen die Daten in einer Region stark konzentriert sind, dazu führen, daß zu viele Sprecher in einer Klasse zusammengefaßt werden, wobei auch solche mit einbezogen werden, für die sich in einer anderen Klasse bessere Ergebnisse ergeben würden. Exemplarisch zeigt sich dies in Tabelle 4.7, in der die Verteilungen von Sprechern in der Teststichprobe nach Klassifikation von *WortSatzIV* mit den beiden Verfahren WARD *linking* und *k-means* gegenübergestellt sind.

Wenn man die einzelnen Merkmalssätze in der Rangliste betrachtet, so fällt auf, daß der Merkmalssatz *gesMerkmVa* in allen Tabellen den ersten Platz einnimmt; er ist jedoch mit Vorsicht zu betrachten, wie die letzte Spalte zeigt: Aufgrund der Fehlermeldungen bei den CMU-Tools ergab sich bei einem beträchtlichen Anteil der Sprecher in der Teststichprobe kein Resultat.

Es stellt sich natürlich die Frage, inwieweit Ergebnisse, die lediglich für einen Teil der Stichprobe vorliegen, noch als relevant bezeichnet werden können. Dies ist sicherlich der Fall, wenn nur 2 von 79 Sprechern fehlen, muß aber verneint werden, wenn die Stichprobe um mehr als 50% schrumpft. Der Perplexitätswert unter 9 scheint außerdem nicht realistisch zu sein, wie die anderen Einträge zeigen, und so liegt denn auch $\wp_x^{ges}$ niedriger als $\wp_x^{s}$. CD 4 zeigt auch hier das bessere Verhalten, obwohl für *gesMerkmVa* noch weniger Sprecherwerte zur Verfügung standen: Zum einen ist die Stichprobenperplexität mit über 9 repräsentativer für die Stichprobe, zum anderen ergibt sich auch eine leichte Reduktion gegenüber dem Gesamtmodell.

Bei Merkmalssatz *WortSatzVa* war die Klassenzahl nach der Clusteranalyse nicht eindeutig festzulegen, und er zeigt auch jetzt noch je Verfahren unterschiedliche Ergebnisse: Für *k-means* sind 5 Klassen, für WARD *linking* hingegen 4 Klassen besser bewertet; die uneinheitlichen Resultate zur Klassenzahl nach der Clusteranalyse bestätigen sich also. Jedoch ist hier insgesamt für die 4 Klassen mit WARD *linking* zu votieren, weil die Stichprobenperplexität von allen vier in Frage kommenden Werten am geringsten ist. Allerdings ist selbst in diesem Fall sogar eine leichte Verschlechterung der Stichprobenperplexität eingetreten, so daß die ursprüngliche Idee bei der Definition dieses Merkmalssatzes[14], nicht von Erfolg gekrönt war. – In gleicher Weise läßt sich *gesMerkmVc* die Entscheidung zugunsten der Klassenzahl 4 und des Verfahrens *k-means* treffen.

[14]nämlich *WortSatzIV* durch ein zusätzliche spontansprachliche Variable zu verbessern

Wenn man die Erkenntnisse aus den Stichprobenperplexitäts-Tabellen kurz und prägnant zusammenfassen möchte, so bleibt festzuhalten:

1. Einfache, naheliegende Merkmalssätze sind nicht unbedingt eine schlechte Wahl (s. die Werte von *WortSatzIIa*).
2. Bigramme als Merkmale *sind* eine schlechte Wahl[15]
3. Spontansprachliche Merkmale bringen relativ gute, aber nicht die besten Ergebnisse (s. *gesMerkmVII*, aber andererseits *WortSatzVa* im Vergleich zu *WortSatzIV*).
4. Insgesamt sind die Ergebnisse bei der Perplexitätsreduktion nicht allzu gut (das Maximum sind 7% für *WortSatzIV* mit einem Bigramm-Modell für CD 4).

[15] Die inhalitche und intuitive Nähe zur Bewertung über n-Gramm-Sprachmodelle hat sich in keiner Weise auf das Ergebnis der Klassifikation ausgewirkt.

Kapitel 5

Interpretation der Klassenstrukturen

5.1 Allgemeine Überlegungen

Es wurde bereits angedeutet, daß eine linguistische Interpretation der am besten bewerteten Merkmalssätze und Partitionen, falls sie überhaupt möglich ist und sinnvolle Ergebnisse hervorbringt, die sprachwissenschaftliche „Krönung“ statistisch noch so hervorragender[1] Ergebnisse darstellt.

Die zu untersuchenden Fragestellungen sind klar:

- Wie lassen sich die statistischen Daten zusammenfassen?
- Welche Sprechertypen sind auszumachen?
- Sind die Sprechertypen linguistisch motiviert und anschaulich erklärbar?

Ihre Beantwortung fällt natürlich umso schwerer, je größer und heterogener der jeweils betrachtete Merkmalssatz ist. Jedoch ist es unumgänglich, wenigstens den Versuch zu unternehmen, die linguistische Relevanz der vorliegenden Ergebnisse zu belegen, v. a. im Hinblick auf die möglichen Einwände von Seiten der Verfechter einer rein qualitativen, nur auf literaturwissenschaftliche Erfahrung und Intuition bauenden Methodik stilistischer Untersuchungen.

Darüber hinaus ist allein schon deshalb, weil keine „natürlichen Klassenstrukturen“ aufgedeckt werden konnten, eine möglichst schlüssige Interpretation derjenigen Partitionen geboten, die mit den am besten bewerteten Merkmalssätzen erzeugt wurden.

Begeben uns also auf das Feld der stets etwas spekulativen linguistischen Interpretation, ohne dabei den sicheren Boden der Statistik völlig zu verlassen. Im Ablauf werden wir wieder auf die Gruppen von Merkmalssätzen rekurrieren, die in 2.5 angegeben wurden.

5.2 Interpretation von Satz- und Turnlängen

Ein Merkmalssatz, der zwei positive Eigenschaften vereint, nämlich eine passable Reduktion der Perplexität zu ermöglichen und relativ gut linguistisch interpretierbar zu sein, ist der aus den Merkmalen ADJ-Wortlängen-Entropie, Satz- und Turnlänge bestehende *WortSatzIV*.

[1] und auch nicht ganz so hervorragender, wie im vorliegenden Fall

Wie bereits aufgrund der geringen Korrelationen zu vermuten war, treten Satz- und Turnlängen in den Ausprägungen kurz–kurz, kurz–lang und lang–kurz unabhängig voneinander auf. Lediglich Sprecher mit gleichermaßen hoher Satz- und Turnlänge finden sich nicht: Es scheint eine Obergrenze für die Anzahl von Wörtern in einem Turn zu geben, die weder mit wenigen langen noch mit vielen kurzen Sätze überschritten wird.

Eine Variable in *WortSatzIV* ist nicht wirklich zu interpretieren: Es handelt sich um die ADJ-Wortlängen-Entropie, auf die wir im folgenden etwas genauer eingehen wollen.

Wenn man sich die Definition der Entropie in Abschnitt 2.2.3, S. 19 vor Augen hält, gelangt man zu dem Schluß[2], daß dieses Merkmal die Diversität der ADJ-Wortlängen in folgender Weise charakterisiert: Je niedrigere Werte sie annimmt, desto ähnlicher sind die relativen Häufigkeiten der einzelnen Wortlängen. Sie wird maximal, wenn alle im Text enthaltenen Wörter (in unserem Fall ADJ) die gleiche Länge haben[3]. Ein hoher Wert der ADJ-Wortlängen-Entropie sagt also aus, daß der betreffende Sprecher überwiegend ADJ ähnlicher Wortlänge verwendet, aber nicht, ob es sich dabei um lange oder kurze ADJ handelt[4].

Man kann natürlich die clusterbezogenen Statistiken in einer Tabelle präsentieren und daraus die für die Separierung der Klassen verantwortlichen Variablen erkennen, aber bei dem aus nur drei Merkmalen bestehenden *WortSatzIV* bietet sich eine grafische Darstellung eher an. Die durch WARD *linking*, das Verfahren mit den besten Stichprobenperplexitätsresultaten für *WortSatzIV*, ermittelten 4 Cluster sind in Abb. 5.1 veranschaulicht.

Aus der Grafik läßt sich entnehmen, daß die Klassen 1 und 2 v. a. durch die ADJ-Wortlängen-Entropie getrennt werden: Beide weisen ähnliche Satz- und Turnlängen im Bereich von 8 – 10 Wörtern/Satz bzw. 1 – 2 Sätzen/Turn auf, jedoch besitzen die Sprecher in Klasse 1 eine niedrigere ADJ-Wortlängen-Entropie als alle anderen. Klasse 3 zeichnet sich durch hohe Satzlängen (> 12) bei ADJ-Wortlängen-Entropie und Turnlängen im Durchschnittsbereich aus, während Klasse 4 im wesentlichen durch die überdurchschnittliche Länge der Turns bei geringer Satzlänge charakterisiert werden kann (wir haben schon darauf hingewiesen, daß die Wortlänge pro Turn nicht sehr variiert).

Versuchen wir noch eine Interpretation der in den einzelnen Klassen versammelten Sprechertypen, wobei wir durchaus in Kauf nehmen, uns zugunsten griffiger Formulierungen einiger Übertreibungen schuldig zu machen.

Klasse 1 enthält im Rahmen dieser Klassifikation den „Durchschnittssprecher“: Keiner seiner Werte ist besonders hoch oder niedrig, also lassen sich auch keine besonderen Aussagen über seine sprachlich-stilistischen Besonderheiten machen. Klasse 1 dient als Referenz für die anderen Klassen.

Die Sprecher in Klasse 2 haben eine eher niedrige ADJ-Wortlängen-Entropie; bis auf die oben angeführten Überlegungen läßt sich allerdings nichts weiter über sie aussagen,weil sie ansonsten dem Durchschnitt in Klasse 1 entsprechen.

Der typische Sprecher aus Klasse 3 formuliert Sätze, die so lang sind, daß nur wenige in einen Turn passen, weshalb die Turnlänge dieser Sprecher auch so kurz ist. Man könnte den Sprechstil dieser Sprecher durchaus mit einigem Recht als „langatmig“ bezeichnen, obwohl eine dialogerhaltende und -befördernde Intention dahinter stecken mag.

Wenn man die Klasse 4 mit einem plakativen Schlagwort umreißen sollte, so wäre vielleicht

[2] vgl. [2], S. 41

[3] Aus der Sicht der Entropie charakterisiert eine Gleichverteilung die höchste Ordnung und den niedrigsten Informationsgehalt.

[4] vgl. 2.3.1; die Höhe der ADJ-Wortlängen-Entropie ist nicht direkt quantitativ interpretierbar

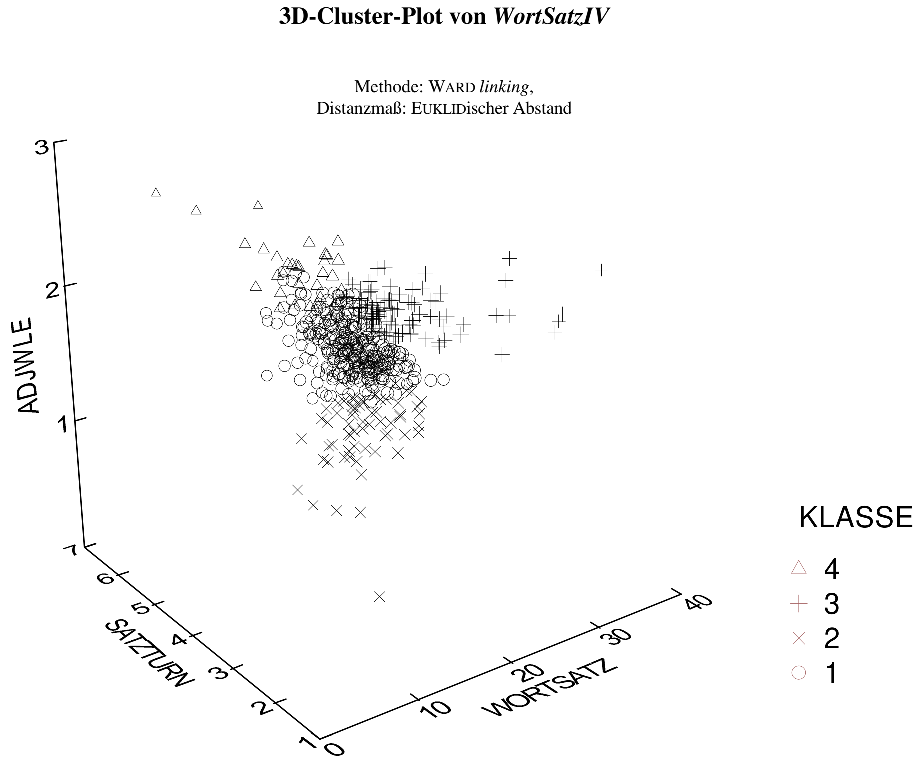

Abbildung 5.1: 3D-Clusterplot von *WortSatzIV*

„gehetzte" Sprecher ein passender Ausdruck: Die hohen Turnlängen bei niedriger Satzlänge bedeuten, daß diese Sprecher stakkatoartig ihre kurzen Äußerungen tätigen, ohne den Gesprächspartner dazwischen allzu oft zu Wort kommen zu lassen — ein für eine Terminverhandlung auch nicht gerade adäquates Verhalten.

Nach diesen mehr negativen Charakterisierungen bleibt jedoch festzuhalten, daß die Klasse der „Durchschnittssprecher", denen diese Eigenschaften nicht zugeschrieben werden können, den deutlich größten Anteil ausmacht; die anderen Klassen dürfen als eher weniger häufige Abweichungen betrachtet werden.

Abschließend können wir folgende kurze Übersicht der Klassen und Sprechertypen geben (Tabelle 5.1).

Wenn wir das zusätzliche Merkmal von *WortSatzVa*, die Aussprachevarianten/Turn, in

Tabelle 5.1: Sprecherklassen von *WortSatzIII*

Klasse	Charakteristik	Interpretation
1	Durchschnitt	„Durchschnittssprecher"
2	niedrige ADJ-Wortlängen-Entropie	„Durchschnittssprecher II"
3	hohe Satzlänge	„langatmiger Sprecher"
4	hohe Turnlänge	"gehetzter Sprecher"

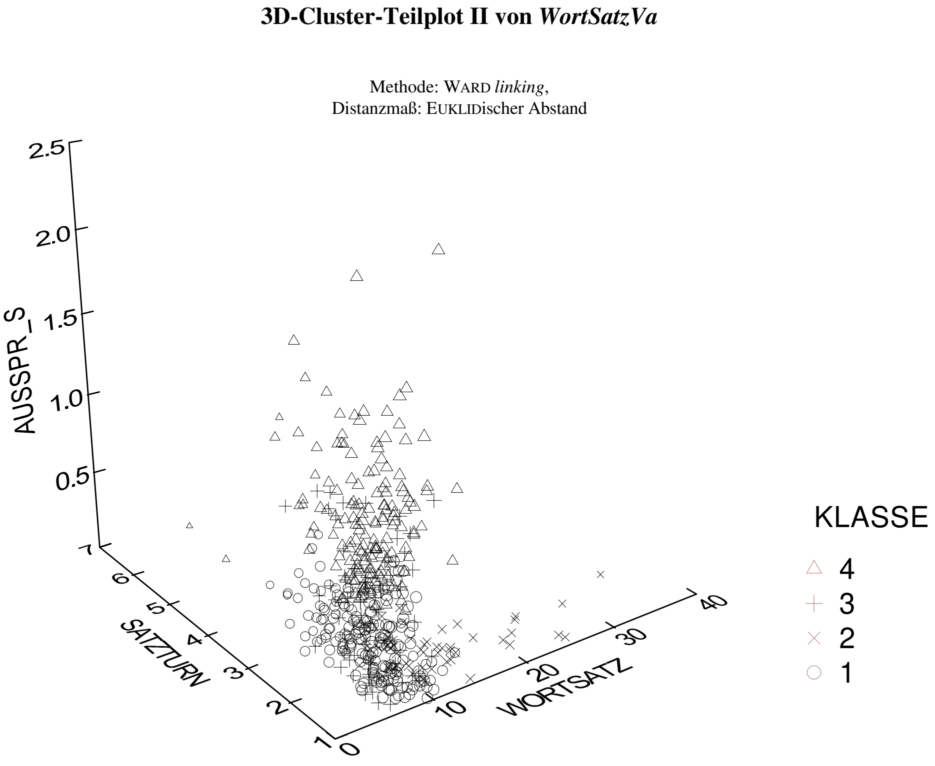

Abbildung 5.2: 3D-Clusterplot II von *WortSatzVa*

dieses Schema mit einbeziehen wollen, dann müssen wir allerdings feststellen, daß sich nicht ohne Grund für *WortSatzVa* schlechtere Resultate bei der Stichprobenperplexität ergeben haben als für *WortSatzIV*, ja sogar niedrigere Bewertungen als bei *WortSatzIIa*.

Die ursprüngliche Begründung für die Wahl der Variablen Aussprachevarianten/Turn war ja, daß die relativ zahlreichen Sprecher ohne Aussprachevarianten als eigene Klasse anzusehen sein müßten. Nun, wenn wir Abb. 5.2 betrachten, so stellen wir fest, daß das Verhältnis genau umgekehrt ist: Die Klassen 1 bis 3 haben so gut wie keine Aussprachevarianten, nur die Sprecher der Klasse 4 benutzen eine ansehnliche Zahl ugs. und dialektaler Wendungen. Die Klassen sind jedoch recht stark durchmischt, so daß sich sogar in Klasse 4 Sprecher fast ohne Aussprachevarianten finden.

Wir können immerhin in einer etwas idealisierten Sichtweise die zwei Klassen der „Durchschnittsprecher" von *WortSatzIV* mit den strukturalistisch notierten distinktiven Merkmalen *+ugs./dialektal* und *-ugs./dialektal* versehen.

5.3 Interpretation bereichsübergreifender Merkmalssätze

Bei den bereichsübergreifenden Merkmalssätzen hat sich nur für *gesMerkmVII* eine Reduktion der Stichprobenperplexität ergeben; wir werden uns also bei der Interpretation nur mit

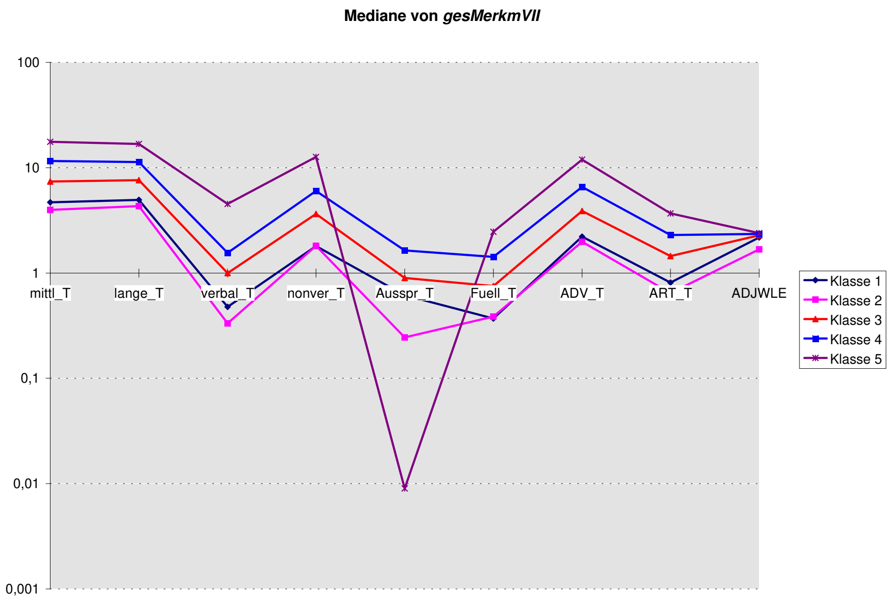

Abbildung 5.3: Cluster-Mediane von *gesMerkmVII*

diesem Merkmalssatz befassen und *gesMerkmVa*, *gesMerkmVb* und *gesMerkmVc* beiseite lassen. Allerdings erweist sich eine Interpretation von *gesMerkmVII* als nicht sehr ertragreich, und wir haben sie nur wegen des relativ guten Stichprobenperplexitätswerts unternommen.

Wegen der Hochdimensionalität des Merkmalsraums muß man auf eine direkte grafische Darstellung verzichten; die einzelnen Klassen lassen sich statt dessen über ihre Mediane als typische Vertreter (Abb. 5.3) beschreiben.

Man erkennt, daß die Wortlängen- und die Kategoriemerkmale monoton mit der Klassennummer zunehmen; das heißt nicht, daß sie nichts zur Differenzierung beitrügen, aber man hätte sich wohl auf ein oder zwei Merkmale aus dieser Gruppe beschränken können,

Die Klassen 1 und 2 werden hauptsächlich uber die ADJ-Wortlängen-Entropie getrennt. Während bei den Mittelwerten der turnbezogenen Variablen — aufgrund der Korrelationen nicht ganz unerwartet — ein fast monotoner Anstieg von Klasse 2 bis 5 zu verzeichnen ist, liegen die Mittelwerte von Klasse 1 und 2 bei diesen Merkmalen nahe beieinander.

Klasse 3 stellt in jeder Hinsicht den Durchschnitt dar, um den sich die Klassen 1 und 4 als unter- bzw. überdurchschnittliche Werte gruppieren. Man muß jedoch beachten, daß es sich nicht um relative Häufigkeiten handelt, sondern um turnbezogenene Werte und deshalb auch eine Korrelation zur Turnlänge gegeben ist. Es ist also nicht ohne weiteres statthaft, die Klassen 3 und 4 als weniger bzw. in höherem Maße spontansprachlich geprägt einzustufen.

An Klasse 5 schließlich ist der extrem niedrige Aussprachevarianten-Wert auffällig, der in scharfem Kontrast zu den anderen hohen Werten steht. Diese Klasse ist die einzige, die eine klare Interpretation zuläßt: Man könnte den Stil iherer Sprecher als „spontan-, aber doch hochsprachlich" charakterisieren.

Auf eine tabellarische Übersicht verzichten wir, weil lediglich für eine einzige Klasse ein lingustisch interpretierbarer Sprechstil ausgemacht werden könnte.

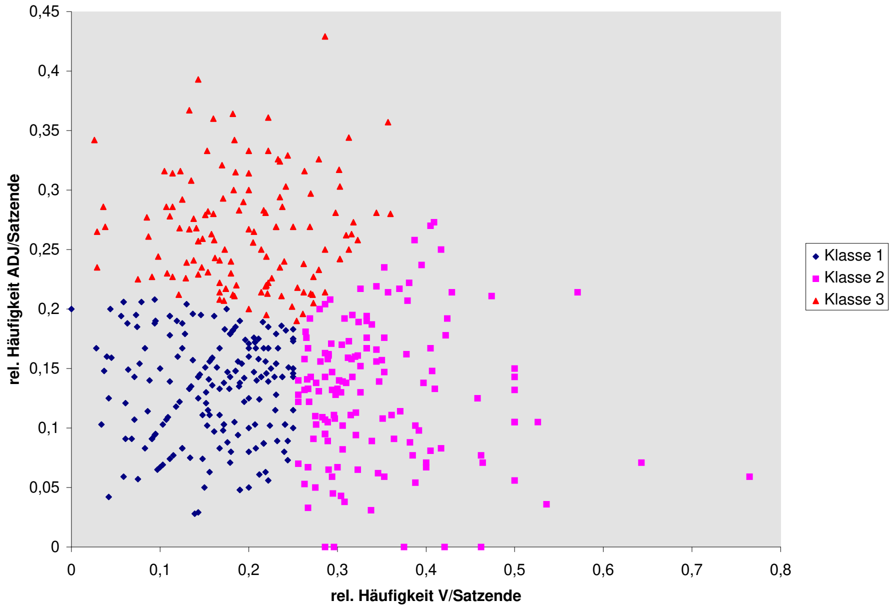

Abbildung 5.4: Clusterplot von *relEnde* mit *k-means*

5.4 Interpretation von *relEnde*

Zur Interpretation von *relEnde* schließlich läßt sich aufgrund der Ergebnisse bei den Perplexitätsuntersuchungen wider Erwarten nicht die visuelle Klassifikation in Abb. 2.7, S. 30 verwenden, sondern es muß auf das Resultat von *k-means* Bezug genommen werden, das wir bequemlichkeitshalber noch einmal in Abb. 5.4, S. 66 dargestellt haben.

Die drei Klassen, die von *k-means*, dem besten bewerteten Verfahren für diesen Satz, gebildet wurden, lassen sich zunächst folgendermaßen quantitativ beschreiben:

In Klasse 1 sind die Häufigkeiten beider Merkmale unter 0,2, sie enthält also Sprecher, die weder viele ADJ noch viele V am Satzende haben. Demgegenüber befinden sich in Klasse 2 Sprecher mit hoher V- ($>= 0,3$) und geringer ADJ-Häufigkeit, während in Klasse 3 die Verhältnisse genau umgekehrt sind (die V-Werte sind klein, die ADJ-Werte $>= 0,4$).

Man kann nun Untersuchungen über eine weitere syntaktische Differenzierung und die semantische Natur der beiden Kategorien am Satzende, die in *relEnde* vertreten sind, anstellen. Finite Verbendstellung ist im Deutschen nur bei Nebensätzen vorhanden[5], die in spontaner Sprache, wo parataktische Kontsruktionen dominieren, in der Minderzahl sind. Daher auch das Verhältnis von 30:70 zu den infiniten V am Satzende, die zu Modalverben im Indikativ oder Konjunktiv als Infinitivergänzungen treten. Sie stellen pragmatisch gesehen meist höfliche Aufforderungen dar bzw. diene der Kennzeichnung von Wünschen[6].

Die ADJ am Satzende sind etweder attributiv oder prädikativ bzw. adverbial gebraucht.

[5] obwohl in der Linguistik die Verbendstellung als Grundtyp gilt, ist sie doch weder in in Affirmativ- noch in Interrogativsätzen realisiert

[6] Das stereotype „Ich würde gerne einen Termin ausmachen" zieht sich aufgrund der Domäne zwangsläufig durch das gesamte VM-Korpus.

Tabelle 5.2: Sprecherklassen von rel. Häufigkeiten am Satzende mit *relEnde*

Klasse	stat.-ling. Beschreibung	Sprechertyp
1	niedrige ADJ- und V-Häufigkeit	"entschlossen"
2	hohe V-Häufigkeit	„höflich„, „umständlich"
3	hohe ADJ-Häufigkeit	„emotional", „gesprächspartnerbezogen"

Im ersten Fall handelt es sich fast ausnahmslos um Datumsangaben, der zweite Fall wird v. a. für Beurteilungen (von Terminvorschlägen) und Zustimmung bzw. Ablehnung[7] verwendet; das Verhältnis attributiver zu prädikativen ADJ beträgt ungefähr 1:2.

Was bedeuten nun diese Erkenntnisse für die Formulierung von Sprechertypen?

Wenn im folgenden eine kurze psycholinguistisch angehauchte Deutung versucht werden soll, so ist diese nur als Darstellung des „idealen Repräsentanten" der jeweiligen Klasse zu sehen und beansprucht keine Allgemeingültigkeit für alle Sprecher dieser Klasse. Für eine unumstößliche Interpretation sind die ermittelten Cluster nämlich nicht separiert genug, wie man Abb. 5.4 entnehmen kann, und es wäre die Hinzuziehung einer ganzen Reihe weiterer Merkmale nötig.

Beginnen wir mit der am einfachsten zu deutenden Klasse 2: Ein relativ großer Anteil an Sätzen Verbendstellung[8] bedeutet eine hohe Frequenz phraseologischer, modalverblicher und/oder konjunktivischer Wendungen. Den Sprechstil einers Sprechers mit dieser Charakterisitk könnte man als „höflich, aber umständlich" bezeichnen.

In Klasse 3 dagegen, in der ADJ-Endstellung dominiert, die sich semantisch überwiegend aus Bewertungen von Aussagen des Gesprächspartners in kurzen Antwortsätzen herleitet, herrscht ein Sprechertypus vor, den man mit einem „starken Partnerbezug" beschreiben könnte. Die naheliegende Hypothese, daß die Häufung von emotional gefärbten ADJ wie „wunderbar„, „gut", und „recht" am betonten Satzende auf eine starken Anteil umgangs- und spontansprachlicher Elemente schließen läßt, m. a. W. daß diese Sprecher sich nicht allzu domänenkonform verhalten[9], kann aber aufgrund fehlender Korrelation der Variablen „ADJ-Häufigkeit am Satzende" zu den spontansprachlichen Merkmalen nicht aufrechterhalten werden.

Eine Interpretation von Klasse 1 ist nicht ganz leicht; die nur mit dem Terminus „Durchschnitt" abzutun, wird der Situation sicherlich nicht gerecht, weil sich diese Sprecher ja gerade durch unterdurchschnittliche Werte in beiden Varaiblen auszeichnen. Vielleicht läßt sich eine Definition aus der Negation herleiten: Ein Sprecher, der weder häufig phraseologische Wendungen benutzt noch sich allzu oft in kurzen Antworten direkt zu Vorschlägen des Partners äußert, dürfte sich seiner Sache sicher sein[10]; wir wollen diese Klasse als die „bestimmten", „entschlossenen" Sprecher kennzeichnen.

Abschließend geben wir wieder eine Übersicht der einzelnen Klassen und ihrer Interpretation an (Tabelle 5.2).

An dieser Stelle wäre eine Zusammenfassung der einzelnen Partition und eine Synopse bei

[7] „frei" (für einen Termin) und „möglich", „gut" und „recht" sind typische Vertreter solcher ADJ
[8] infiniter Verben, man erinnere sich an das beschriebene Verhältnis von 30:70
[9] in einer Terminverhandlung ist ja vor allem Sachlichkeit gefragt
[10] im Gegensatz zum Autor bei dieser Interpretation

der Interpretation wohl wünschenswert. Wir müssen diese aufwendige Arbeit jedoch aufgrund der Heterogenität der Partitionen und des immer begrenzten Rahmens einer solchen Arbeit weiterführenden linguistischen Spezialuntersuchungen überlassen.

Kapitel 6

Anwendung und Ausblick

6.1 Ansätze zur Verwendung in der Sprachverarbeitung

Nach der Definition der Merkmale, Bewertung der Partitionen und ihrer Interpretation stellt sich nun die Frage, wie die ermittelten Stilklassen und die daraus gewonnenen stilspezifischen Sprachmodelle in einem Sprachverarbeitungssystem nutzbar gemacht werden können. Ohne dabei die Implementation in ein konkretes System auszuführen, sollen im folgenden einige Prinzipien und Möglichkeiten für eine solche Umsetzung angesprochen werden, die insbesondere für den in dieser Arbeit nicht unbedingt eingetretenen Fall guter Ergebnisse bei der Perplexitätsreduktion anwendbar sind.

Aus Sicht der Praxis bei der Implementierung darf nicht übersehen werden, daß die Ermittlung von Merkmalen während des Sprechens[1] und die darauf gestützte Auswahl eines geeigneteren Sprachmodells auch Rechenzeit in Anspruch nehmen, die man mit dem evtl. für den aktuellen Sprecher weniger gut geeigneten globalen Modell für die Untersuchung weiterer Worthypothesen nützen könnte. Mit anderen Worten, der verringerte Umfang an Verzweigungsmöglichkeiten im Baum der Worthypothesen, der durch ein sprecheradäquates Wortmodell erreicht werden kann, darf nicht von der für seine Etablierung erforderlichen Rechenzeit aufgezehrt werden.

Ein erster Ansatz besteht in der Idee von Abschnitt 4.3.2, die Zuordnung statt über die Merkmale direkt über den minimalen Perplexitätswert vorzunehmen, wie in 4.3 angedeutet. Die Perplexitätsberechnung mit allen in Frage kommenden Werten ist noch aufwendiger als die Bestimmung von Merkmalen, so daß, was eben für Merkmale angedeutet wurde, in noch stärkerem Maß gilt: Die zusätzliche Rechenzeit wird den Spezialisierungsgewinn zunichte machen.

Grundsätzlich muß darauf geachtet werden, daß die erforderlichen Merkmale möglichst frühzeitig und zuverlässig ermittelt werden können. Dabei eignen sich Merkmale auf Wortebene generell besser als solche auf Turnebene; relative Häufigkeiten sind ebenfalls erst ab größeren Datenmengen aussagekräftig. Andererseits haben sich aber in dieser Arbeit gerade Intra-Wort-Merkmale als nicht brauchbar erwiesen (vgl. die Ergebnisse bei der Merkmalsauswahl und die Perplexitätsresultate bei den Partitionen mit Wortlängenmerkmalen). Die ADJ-Wortlängen-Entropie ist zwar ein Wortlängenmerkmal, aber über die Verteilung defi-

[1] Denn die Spracherkennung der jeweils folgenden Äußerungen kann natürlich nur dann verbessert werden, wenn wirklich noch weitere Äußerungen folgen; a posteriori-Betrachtungen sind lediglich von theoretischem Interesse.

niert und ebenfalls erst ab einer größeren Textmenge sinnvoll. In der Mitte zwischen wort- und satz- bzw. turnbezogenen Werten liegen Kategorienhäufigkeiten; hier haben sich v. a. für die ADV-Verteilung noch die besten Resultate ergeben (vgl. die Werte von *gesMerkmVa*und *gesMerkmVII*). Idealerweise sind Kategorienwerte aber auch erst nach einer Anzahl von Sätzen zu ermitteln, so daß man wohl doch auf die Satz- und Turnlängen angewiesen ist, für deren Ermittlung einiges an Sprecherdaten erforderlich ist.

Der Aspekt der inkrementellen Verarbeitung kommt aber nicht nur in den besseren Klassifikationen durch höhere Datenfülle zum Tragen, sondern auch in dem anfänglichen Verzicht auf kompliziertere Merkmale, auch wenn ihre Berechnung mit der verfügbaren Datenmenge bereits möglich wäre. Dies war ja gerade die Motivation zur Definition von und zur Klassifikation mit *WortSatzIIa*, und die Ergebnisse für diesen Merkmalssatz sind ja in der Tat nicht wesentlich schlechter als für *WortSatzIV*. Stellt sich keine Verbesserung bei der Erkennungsleistung ein, so kann man unter Verwendung weiterer Merkmale die Klassifikation verfeinern oder sogar revidieren.

Neben der schnellen Merkmalsextraktion zur Einordnung unbekannter Srprecher in die Stilklassen ist aber auch die Frage von Bedeutung, inwieweit sich das Problem der ungesehenden Wortfolgen[2] durch eine Kombination von Stilklassenmodellen mit dem allgemeinen Modell angehen läßt. Dies gilt umso mehr, als die Verbesserung eines Sprachmodells durch seine Spezialisierung mit Hilfe der Stilklassen aufgehoben und sogar konterkariert werden kann — durch die einfache Tatsache, daß die spezielle Modelle stets weniger Daten zur Verfügung hatten als das allgemeine, und daß demnach bei einem neuen Sprecher den Stilklassenmodellen stets mehr n-Gramme unbekannt sein werden als dem Gesamtmodell.

Für diese Problematik lassen sich die Verfahren einsetzen, die zur Glättung von Verteilungen, d. h. zur Aufteilung der verfügbaren Wahrscheinlichkeitsmasse in einen Bereich für bekannte und einen Bereich für unbekannte Wörter, entwickelt wurden[3]. Man unterscheidet dabei zwei grundsätzliche Methoden, nämlich die Rückfall- und die Interpolationsstrategie.

Bei der Interpolationsstrategie werden mehrere Modelle nebeneinander benutzt, indem die Wahrscheinlichkeit einer Wortfolge durch Summation über die Wahrscheinlichkeiten der verwendeten Modelle berechnet wird; die Wahrscheinlichkeiten sind dabei so zu gewichten, daß sich wieder eine Wahrscheinlichkeitsverteilung über die Wortfolgen ergibt, wobei ein Optimum aus der Kombination der Wahrscheinlichkeiten der beteiligten Modelle gesucht wird.

Bei der Rückfallstrategie hingegen wird zur Berechnung der Wahrscheinlichkeit einer Wortfolge stets genau ein Modell verwendet, nämlich bei bereits gesehenen Wortfolgen das normale Modell und bei einer unbekannten Wortfolge das Rückfallmodell; der Rückfall (engl. *backing off*) erfolgt bei einer ungesehenen Wortfolge vom n-Gramm-Modell auf das (n-1)-Gramm-Modell, also bis hinunter zu Unigrammen.

Diese Strategie kann man nun auf das Paar spezielles Stilklassenmodell/allgemeines Sprachmodell übertragen: Bei einem neuen Sprecher muß ohnehin zunächst das allgemeine Modell verwendet werden. Ist die Zuordnung zu einer Stilklasse klar, so wird auf das entsprechende Klassenmodell gewechselt; wenn aber dem spezifischen Modell unbekannte Wortfolgen auftauchen, wird dafür wieder das allgemeine Modell benutzt.

Ein Ansatz, der nicht direkt mit den Stilklassen in Verbindung steht, sondern eher als orthogonales Verfahren zur Sprecheradaption von Sprachmodellen angesehen werden kann, ist das in [25] vorgestellte *caching*, ein Verfahren, das — nomen est omen — die jeweils letzten

[2] s. die in 4.3 gegebenen Erläuterungen

[3] zu diesen Verfahren vgl. wiederum die Ausführungen in [22] und [23]

Worte eines Sprechers in einem „Kurzzeitgedächtnis“ hält und damit bereits eine Art Stilmodellierung vornimmit (wenn sich nämlich ein Sprecher durch typische, kurze Wortfolgen chearkterisieren läßt). Dieses Verfahren wird jedoch in [16] mit einem Rückfallmodell kombiniert, das auf der Methode des *absolute discounting*[4] beruht, so daß sich eine Möglichkeit ergibt, Stilklassen, *caching* und *backing off* zu verbinden und damit die Perplexität weiter zu reduzieren.

Eine Modifikation der Rückfallstrategie, *language fill-up*, wird in [4] zur Sprecheradaption in einem Diktiersystem eingesetzt. Dabei wird ein sprecherspezifisches Modell in Kombination mit einem allgemeinen, feststehenden Modell verwendet. Zunächst wird ersteres auf sprecherspezifischen Daten trainiert, und den Wahrscheinlichkeiten wird nach der Methode *absolute discounting* Wahrscheinlichkeitsmasse für ungesehene Wortfolgen (die Autoren beschränken sich auf Bigramme) entzogen. Ist das Bigramm dann im Verlauf des Diktats dem sprecherspezifischen Wortmodell bekannt, wird die Wahrscheinlichkeit daraus unverändert übernommen. Wenn das Bigramm dem allgemeinen Modell bekannt ist, wird das Sprachmodell mit dem Wert aus dem allgemeinen Modell „aufgefüllt“; falls nicht, erfolgt der Rückfall auf die Unigramm-Wahrscheinlichkeit des speziellen Modells. — Dieses Verfahren läßt sich in gleicher Weise für Stilklassen anwenden: Das klassenspezifische Modell übernimmt dabei die Rolle des spracheradaptierten Modells im *language fill-up*, das allgemeine Sprachmodell den Part des globalen Modells.

Ein anderer Formalismus für Sprachmodelle sind stochastische Grammatiken[5], also Grammatiken, bei denen der Symbolübergang bei den Produktionsregeln mit Wahrscheinlichkeiten belegt ist (über jeder nichtterminalen Kategorie ist dabei demnach eine Wahrscheinlichkeitsverteilung definiert). Dennoch lassen sich die in der Partitionierung der Sprecher in Stilklassen enthaltenen Informationen auch für stochastischen Grammatiken nützen: Unter der Annahme, daß sich die ermittelten Sprechstilklassen auch in grammatischen Spezifika zeigen[6], lassen sich die Übergangswahrscheinlichkeiten mit dem Textmaterial aus den einzelnen Stilklassen trainieren und so separate, klassenspezifische Grammatiken ermitteln. Dies geschieht natürlich in der Hoffnung, daß bisher unbekannten Sprechern nach Einordnung in ihre Stilklasse eine besser auf ihren Sprachgebrauch adaptierte Grammatik zugeordnet werden kann.

Zum Abschluß dieser Überlegungen zum Einsatz von Stilklassen in Sprachverabeitungssystemen soll noch eine Verbindung zu dem in [13] beschriebenen Ansatz hergestellt werden. Die Autoren betrachten in diesem Beitrag eine Kombination aus einem linguistischen und einem stochastischen Sprachmodell. Dabei werden beide Modelle durch Kategorien-Bigramme repräsentiert und zu einem parallelen Modell verschmolzen, auf dem dann Standard-Suchverfahren operieren können. Dabei soll das linguistische, spezialisierte Modell die grammatisch korrekten Sätze verbessern, während das stochastische, robustere Modell die spontansprachlichen Aspekte der Spracherkennung besser verarbeiten kann. Dieser Ansatz wird dann auf turnbasierte Modelle angewendet. — Nicht nur, daß auf einer abstrakteren Ebene Stilklassen, die zur Verfeinerung von Sprachmodellen verwendet werden, ebenfalls als eine solche

[4] d. h. die verfügbaare Wahrscheinlichkeitsmasse, die für ungesehene Wortfolgen reserviert ist, wird von den aus den relativen Häufigkeiten der einzelnen Wortfolgen ermittelten Schätzwerten durch Subtraktion eines konstanten Wertes gewonnen — im Unterschied zum *relative discounting*, wo diese Subtraktion proportional zur relativen Häufigkeit erfolgt

[5] vgl. z. B. [35], S. 204 ff.

[6] eine Annahm, die zwar intuitiv gerechtfertigt erscheint, aber statistisch aufgrund geringer Datenmengen pro Sprecher im VM-Korpus schwer nachweisen läßt, was bei allen makrostilistischen Merkmalen der Fall ist; grammatische Merkmale wurden ja in der vorliegenden Arbeit aus diesen theoretischen und aus den in 2.1.2 angegebenen praktischen Gründen nicht untersucht

Kombination von Linguistik und Stastistik betrachtet werden können, sondern auch in einem konkreteren Sinn besteht die Möglichkeit, Wortmodelle sprechstilspezifischer Klassen, die mit spontansprachlichen Merkmalen gebildet wurden (man denke an *gesMerkmVII*), als die stochastische Komponente der Modellkombination von [13] einzusetzen, in der Hoffnung, daß derartige spontansprachlich spezialisierte Stilklassen-Wortmodelle die Aufgabe des stochastischen Teils des Modellpaars noch besser erfüllen können[7].

6.2 Was bleibt zu tun?

Bei näherer Betrachtung stellt sich aber nach den eben geschilderten Einsatzmöglichkeiten von Stilklassen angesichts der Ergebnisse dieser Arbeit doch die Frage, ob bei allen Verbesserungsmöglichkeiten der Perplexitätswerte durch die Auswahl eines sprechstilklassenspezifischen Wortmodells nicht doch die schiere Quantität der Daten zu bevorzugen ist, m. a. W. ob nicht stets alle verfügbaren Daten herangezogen werden müssen und man im Vertrauen auf das Gesetz der großen Zahlen auf eine qualitative, d. h. sprecherangepaßte Verbesserung verzichten sollte. Leider kann man ja nicht damit rechnen, von den einzelnen Klassen genügend Daten in der Trainingsstichprobe zur Verfügung zu haben, und so ist die Versuchung stets gegeben, nur die gesamte Datenmenge zur Erstellung eines Wortmodells zu nutzen.

Wenn wir nun abschließend im Hinblick auf die anfangs aufgestellten Arbeitshypothesen Bilanz ziehen, so muß konstatiert werden, daß die Aufdeckung von möglicherweise in den Daten des VM-Korpus enthaltenen „natürlichen Klassenstrukturen“ nicht gelungen ist, sondern nur „interpretierbare Zerlegungen“ ([40], S. 45) gefunden wurden. Darüber hinaus wurde lediglich bei einigen untersuchten Merkmalen und Partitionen eine Reduktion der Perplexität gegenüber dem Gesamtmodell erreicht, die jeweils recht moderat ausfiel. Die Klassenzuordnung ist, wie die minimalen Perplexitätswerte zeigen, auch nicht wesentlich zu verbessern, so daß die Partitionen als solche, und nicht etwa der eingesetzte 3NN-Klassifikator, für die marginalen Resultate verantwortlich sind.

Was die Perplexitätsreduktion anbelangt, so wurde bereits weiter oben als Begründung angemerkt, daß die Vorteile der Spezialisierung durch die geringere Datenmenge gegenüber dem allgemeinen Modell verloren gehen können: Dieses Argument ist jedoch nicht allzu stichhaltig, weil die Klassenstruktur trotz ihrer linguistischen Interpretierberkeit nicht genügend ausgeprägt ist.

An dieser Stelle ist die Möglichkeit in Betracht ziehen, daß die „Arbeitshypothesen“ nicht zutreffen: Dies kann nur bedeuten, daß sich persönlicher Stil eben nicht quantitativ nachvollziehbar in spontansprachlichen Äußerungen manifestiert, wenigstens nicht in dem in mancher Hinsicht restringierten Bereich des VM-Korpus; eine Auffassung, die zwar der alltäglichen Anschauung widerspricht, aber nach den Ergebnissen dieser Arbeit dennoch nicht ohne weiteres von der Hand zu weisen ist.

Als Begründungsansätze lassen sich auch einige in der eingangs angeführten Untersuchung [14] vorgetragene Charakteristika interaktiven Sprachgebrauchs anführen: Zum einen besteht die Möglichkeit, daß die Sprecher in der Gesprächssituation des VM-Korpus die „situationsgerichteten Normen und Erwartungen so weit wie möglich erfüllen“ ([14], S. 122) und damit unmarkierter Stil vorherrschend ist. Dies wäre mit einer erhöhten Kooperativität der Gesprächsteilnehmer angesichts der Aufnahmesituation zu erklären. Darüber hinaus ist

[7] Es besteht noch eine weitere Gemeinsamkeit zu dem Ansatz von [13]: Hier wie dort sind die Resultate nicht ganz so positiv, wie anfängliche Überlegungen und Untersuchungen in Teilbereichen vermuten ließen.

nach Franck „die Orientierung am jeweiligen Hörer [...] das wichtigste Prinzip interaktiven Sprachgebrauchs“ ([14], S. 128). Dies läßt sich dahingehend interpretieren, daß sprecherspezifische Stilelemente im Dialog gegenüber dem *recipient design* zurücktreten.

Die Konsequenz daraus wäre die Konzentration auf völlig andere Ansätze zur Perplexitätsreduktion und damit zur Leistungsverbesserung von Sprachverarbeitungssystemen. Andererseits geben die Ergebnisse der Arbeit, so unzureichend sie auch sein mögen, durchaus Anlaß zu weiteren quantitativen Untersuchungen zur Klassifikation persönlicher Sprechstile. Wenn man die eingangs aufgestellten Hypothesen akzeptiert, ist im Grunde vollkommen klar, worauf solche Untersuchungen abzielen müssen: Zunächst einmal auf die Sammlung und Auswertung weiteren Datenmaterials, um nicht so sehr die Anzahl der Sprecher zu erhöhen als vielmehr die Textgröße der einzelnen Sprecher.

Der Hauptpunkt aber ist die Ermittlung eines besseren Merkmalssatzes, als es in dieser Arbeit gelungen ist, weil dann erst eine echte Reduktion der Perplexität gegenüber dem allgemeinen Modell erreicht werden kann. Das kann möglicherweise durch eine veränderte Auswahl und Gewichtung aus den bereits untersuchten Variablen geschehen, wenngleich diese Möglichkeit weniger wahrscheinlich ist als diejenige, die sich durch Hinzuziehen von Merkmalen aus ganz anderen Bereichen ergibt.

Ich denke dabei an folgende Bereiche:

Zum einen dürfte sich die Inspektion des Sprachsignals als lohnend erweisen, um Prosodie, Artikulation und Intonation zu extrahieren und näher zu untersuchen.

Zum zweiten könnten in Anlehnung an [39], S. 213 Merkmale berücksichtigt werden, die über den intratextuellen Bereich hinausgehen. Dazu gehören solche aus dem Bereich der Soziolinguistik, die im VM-Korpus (demnächst in auswertbarer und vollständiger Form) als Zusatzinformationen zur Verfügung stehen, etwa Herkunft, Geschlechtszugehörigkeit oder Ausbildung der Sprecher. Es ist beispielsweise durchaus denkbar (vgl. [19]), daß sich geschlechtsspezifische Sprechstile auch quantitativ faßbar ermitteln lassen. — Diese soziolinguistischen Variablen wären im praktischen Einsatz bei einem unbekannten Spracher vor Dialogbeginn zu ermitteln und dem System bekannt zu machen.

Eine linguistisch interessante, wenngleich quantitativ schwer faßbare und in der statistischen Auswertung problematische Möglichkeit zur Auffindung neuer Merkmale ergibt sich aus dem Bereich Makrostilistik. Insesondere in der Pragmatik[8] und Dialogführung[9] lassen sich Untersuchungen anstellen. Allerdings sind hier völlig neue quantitativ-stilistische Merkmale zu entwickeln, die ja dann stets nur partnerspezifisch zu sehen und zu untersuchen sind.

Wie auch immer, es bleibt zu hoffen, daß weiteren Untersuchungen in diesem Bereich mehr Erfolg beschieden sein wird.

---

[8] wir sind auf Probleme mit dem VM-Korpus in diesem Bereich bereits weiter oben im Abschnitt 2.1.2, S. 8 kurz eingegangen, wo sich bedauerlicherweise die Notwendigkeit eines Verzichts auf pragmatische Merkmale für diese Arbeit ergab

[9] in Anlehnung an [33]

Literaturverzeichnis

[1] **Altmann, Gabriel**. *Status und Ziele der quantitativen Sprachwissenschaft.* In: **Jäger, Siegfried** (Hg.). *Linguistik und Statistik.* Braunschweig 1972, S. 1 – 11

[2] **Altmann, Gabriel**. *Wiederholungen in Texten.* Bochum 1988

[3] **Altmann, Gabriel**. *Status und Funktion quantitativer Verfahren in der Computerlinguistik.* In: **Batori, Istvan** (Hg.). *Handbücher zur Sprach- und Kommunikationswissenschaft; Band 4: Computerlinguistik.* New York 1989, S. 113 – 119

[4] **Besling, Stefan; Meier, Hans-Günter**. LANGUAGE MODEL SPEAKER ADAPTION. In: *Europan Conference on Speech Communication and Technology.* Madrid 1995, S. 1755 – 1758

[5] **Batori, Istvan** (Hg.). *Handbücher zur Sprach- und Kommunikationswissenschaft; Band 4: Computerlinguistik.* Berlin/New York 1989

[6] **Bock, Hans Herrmann**. *Automatische Klassifikation.* Göttingen 1974

[7] **Bolz, Norbert**. *Gewinnung und Auswertung quantitativer Merkmale in der statistischen Stilforschung.* In: **Spillner, Bernd** (Hg.). *Methoden der Stilanalyse.* Tübingen 1984, S. 193 – 222

[8] **Buch, Kai Rander**. *A Note on Sentence-Length as Random Variable.* In: **Doležel, Lubomir; Bailey, Richard W.** (Hg.). *Statistics and Style.* New York 1969, S. 76 – 79

[9] **Büning, Herbert; Trenkler, Götz**. *Nichtparametrische statistische Methoden.* Berlin 1978

[10] **Deichsel, Guntram; Trampisch, Hans Joachim**. *Clusteranalyse und Diskriminanzanalyse.* Stuttgart 1985

[11] **Doležel, Lubomir; Bailey, Richard W.** (Hg.). *Statistics and Style.* New York 1969

[12] **Doležel, Lubomir**. *A Framework for the Statistical Analysis of Style.* In: **Doležel, Lubomir; Bailey, Richard W.** (Hg.). *Statistics and Style.* New York 1969, S. 10 – 28

[13] **Eckert, Wieland; Gallwitz, Florian; Niemann, Heinrich**. *Combining Stochastic and Linguistic Language Models for Recognition of Spontaneous Speech.* In: *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing. Atlanta 1996, S.* 423 – 426 Atlanta 1996

[14] **Franck, Dorothea**. *Stil und Interaktion*. In: **Spillner, Bernd** (Hg.). *Methoden der Stilanalyse*. Tübingen 1984, S. 121 – 137

[15] **Gauger, Hans Martin**. *Über Sprache und Stil*. München 1995

[16] **Generet, M.; Ney, H.; Wesset, F.** *Extensions of Absolute Discounting for Language Modeling*. In: *Europan Conference on Speech Communication and Technology*. Madrid 1995, S. 1245 – 1248

[17] **Grotjahn, Rüdiger**. *Linguistische und statistische Methoden in Metrik und Textwissenschaft*. Bochum 1979

[18] **Gupta, V.; Lenning, M.; Mermelstein, P.** *A Language Model for very large Vocabulary Speech Recognition*. In: *Computer Speech & Language, Band 6*. London 1992, S. 331 – 344

[19] **Hiller, Ulrich**. *Some Sex-Related Differences In The Use Of Spontaneous Speech*. Duisburg 1985

[20] **Jekat, Susanne et al.** *Dialogakte in VERBMOBIL*. Verbmobil Technisches Dokument Nr. 26, 1995

[21] **Jäger, Siegfried** (Hg.). *Linguistik und Statistik*. Braunschweig 1972

[22] **Jelinek, Fred**. *Self-Organized Language Modeling for Speech Recognition*. In: **Waibel, Alex; Lee, K.** (Hg.). *Readings in Speech Recognition*. San Mateo, CA. 1990, S. 450 – 506

[23] **Katz, Slava M.** *Estimation of Probabilities from Sparse Data for the Language Component of a Speech Recognizer*. In: *IEEE Transactions on Acoustics, Speech and Signal Processing, Bd. 35, Nr. 3*. New York 1987, S. 400 – 401

[24] **Kohler, Klaus** et al. *Handbuch zur Datenaufnahme und Transliteration in TP14 von VERBMOBIL 3.0*. Verbmobil Technisches Dokument Nr. 11, 1994

[25] **Kuhn, Roland; de Mori, Renato**. *A Cache-Based Natural Language Model for Speech Recognition*. In: *IEEE transactions on pattern analysis and machine intelligence*. New York 1990, S. 570 – 583

[26] **Meier, Helmut**. *Deutsche Sprachstatistik*. Hildesheim Olms 1964

[27] **Metzeltin, Michael**. *Verfahren makrostilistischer Textanalyse*. In: **Spillner, Bernd** (Hg.). *Methoden der Stilanalyse*. Tübingen 1984, S. 49 – 99

[28] **Mucha, Hans Joachim**. *Clusteranalyse mit Mikrocomputern*. Berlin 1992

[29] **Muller, Charles**. *Einführung in die Sprachstatistik*. München 1972

[30] **Pieper, Ursula**. *Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse*. Tübingen 1979

[31] **Popovici, Cosmin; Baggia, Paolo**. *Specialized Language Models Using Dialogue Predictions*. In: *Computational Linguistics*. New York 1996

[32] **Sanders, Willy**. *Linguistische Stiltheorie.* Göttingen 1977

[33] **Sandig, Barbara**. *Ziele und Methoden einer pragmatischen Stilistik.* In: **Spillner, Bernd** (Hg.). *Methoden der Stilanalyse.* Tübingen 1984, S. 137 – 161

[34] **Schiller, A.; Teufel, S.; Thielen, C.** *VORLÄUFIGE guidelines für das Tagging deutscher Textkorpora mit STTS.* Tübingen 1995

[35] **Schukat-Talamazzini, Ernst Günter**. *Automatische Spracherkennung.* Braunschweig 1995

[36] **Sowinski, Bernhard**. *Makrostilistische und mikrostilistische Textanalyse: Thomas Manns 'Luischen' als Beispiel.* In: **Spillner, Bernd** (Hg.). *Methoden der Stilanalyse.* Tübingen 1984, S. 21 – 47

[37] **Spillner, Bernd** (Hg.). *Methoden der Stilanalyse.* Tübingen 1984

[38] **Spillner, Bernd**. *Grundlagen der Phonostilistik und Phonoästhetik.* In: **Spillner, Bernd** (Hg.). *Methoden der Stilanalyse.* Tübingen 1984

[39] **Spillner, Bernd**. *Methoden der Stilanalyse: Forschungsstand und analytische Bibliographie.* In: **Spillner, Bernd** (Hg.). *Methoden der Stilanalyse.* Tübingen 1984, S. 223 – 240

[40] **Stahl, Herbert**. *Clusteranalyse großer Objektmengen mit problemorientierten Distanzmaßen.* Thun 1985

[41] **Waibel, Alex; Lee, K.** (Hg.). *Readings in Speech Recognition.* San Mateo, CA. 1990

[42] **Ward, Wayne; Issar, Sunil**. *A Class Based Language Model for Speech Recognition.* In: *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing.* Atlanta 1996, S. 416 – 418

[43] **Winter, Werner**. *Styles as Dialects.* In: **Doležel, Lubomir; Bailey, Richard W.** (Hg.). *Statistics and Style.* New York 1969, S. 3 – 9

[44] **Zimmer, Rudolf**. *Stilanalyse.* Tübingen 1978