Quantifizierende Textanalyse: mit der Hilfe des Computers auf der Suche nach dem anonymen Autor

Trauth, Michael

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Historical Social Research, Vol. 17 — 1992 — No. 1, 133-141

menarbeit mit Heinrich P. Delfosse und Heinz Schay. Stuttgart-Bad Cannstatt 1986 [FMDA Abt. III, Bd. 5]. XLII, 584 S. (Es sind weitere Bände erschienen oder in Vorbereitung.)

Ders.: *Kant-Index*. Bd. 14: *Personenindex zum Logikcorpus*. Erstellt in Zusammenarbeit mit Heinrich P. Delfosse und Elfriede Reinardt. Stuttgart-Bad Cannstatt 1990 (im Druck).

Ders.: *Die Datierung der Reflexion 3716 und die generellen Datierungsprobleme des Kantischen Nachlasses. Erwiderung auf Josef Schmucker.* In: Kant-Studien 68 (1977) S. 321–340.

Ders.: *Kant per Computer. Einsatzmöglichkeiten der elektronischen Datenverarbeitung im Bereich der Geisteswissenschaften.* In: Neue Deutsche Hefte 1 (1987) S. 106–113.

Ders.: *Reimarus zwischen Wolff und Kant. Zur Quellen- und Wirkungsgeschichte der ›Vernunftlehre‹ von Hermann Samuel Reimarus.* In: Logik im Zeitalter der Aufklärung. Studien zur ›Vernunftlehre‹ von Hermann Samuel Reimarus, hrsg. von Wolfgang Walter und Ludwig Borinski (Veröffentlichung der Joachim Jungius-Gesellschaft der Wissenschaften Hamburg, Nr. 38), Göttingen 1980, S. 9–32.

Quantifizierende Textanalyse. Mit der Hilfe des Computers auf der Suche nach dem anonymen Autor*

Michael Trauth (Trier)

I.

Die Textanalyse blickt auf eine altehrwürdige, bis in die Anfänge der Kulturgeschichte zurückreichende Tradition zurück. Diese Tradition hat jedoch bis heute noch kein kanonisches Verständnis von >Textanalyse<, keine *communis opinio* darüber, was >Textanalyse< sei und wie sie betrieben werden müsse, hervorgebracht. Nicht nur die im engeren Sinne mit Sprache befaßten Wissenschaften reklamieren textanalytisches Arbeiten für sich, sondern auch viele andere wie Theologie, Philosophie, Geschichtswissenschaft, Pädagogik, Psychologie, Jurisprudenz und Kriminologie, um nur die wichtigsten zu nennen. Unter ihnen haben indessen

* Protokoll des 49. Kolloquiums über die Anwendung der EDV in den Geisteswissenschaften an der Universität Tübingen am 7. Juli 1990.

nicht einmal die Einzeldisziplinen ein monolithisches Konzept der Zergliederung von Texten zu entwickeln vermocht; entsprechend disparat präsentiert sich das Gesamtbild. Den vornehmsten Grund für die Heterogenität offenbart ein genauerer Blick auf den Begriff >Textanalyse<: Dessen Etymologie verweist auf die >Auflösung eines Gewebes<, das seinerseits ein vielschichtiges Konstrukt gänzlich verschiedener Bestandteile sein kann: von Zeichen - man beachte die Multivalenz dieses Begriffs! -, von Wörtern, Syntagmen, grammatischen Strukturen, logischen Beziehungen, Semata, Sinnelementen etc. Vielgestaltig war also seit jeher nicht erst die *Methodologie* textanalytischen Arbeitens, sondern bereits das, was analysiert werden sollte.

Die quantitativ-deskriptive Analyse, von der hier gehandelt werden soll, verkörpert nur einen kleinen Ausschnitt aus dem weiten Feld der Textanalyse. Sie ist nicht ganz so alt und ursprünglich wie ihre interpretierende, subjektive, auch >intuitiv< genannte Schwester, geht aber wie diese auf antike Wurzeln zurück. Ihre Besonderheit ist die Verwendung der Mathematik, mit der von alters her der Anspruch auf *intersubjektive Überprüfbarkeit* verknüpft ist. Daß dieser letztere freilich ein ums andere Mal auf der Strecke blieb, als die Quantifizierung auf mitunter groteske Weise zum Vehikel inhaltlicher Inferenz gepreßt wurde, sei hier nur am Rande vermerkt.

Neben dem modernen linguistischen Interesse, Mechanismen, Stand und Entwicklungstendenzen einer Sprache zu beschreiben, stand - und steht - *literarische Kriminalistik* im Vordergrund textanalytischen Fragens. Die Literaturgeschichte kennt zahlreiche Exempla ungeklärter Autorschaften und Datierungen, Probleme, für die beim Versagen anderer Mittel regelmäßig die Quantifizierung der strittigen Dokumente als Remedium herangezogen wurde. Indessen kann auch eine Bilanz solcher Bemühungen nur festhalten, daß die einschlägig tätigen Wissenschaften keinen Konsens über das zweckmäßige Objekt, geschweige denn über die Methode der Quantifizierung erzielt haben.

Hinter dem Einsatz der quantifizierenden Analyse steht naturgemäß die Überzeugung, daß sich die Individualität eines Schreibenden als >Regelhaftes< im Geschriebenen wiederfinden lasse: Vorlieben und Abneigungen etwa im aktiven Wortschatz und in der Wortbildung, in der Fügung von Junkturen, in der Konstruktion von Hypotaxen, in der Einleitung von Sätzen und Abschnitten, überhaupt in der Untergliederung von Texten u. v. a. bis hin zum zyklischen Wiederkehren der gleichen Formulierung desselben Gedankens. Diese nur scheinbar triviale These ist - wiewohl sie nicht ernsthaft bestritten werden kann - zugleich auch die Crux des Ganzen, hat sie doch immer wieder zur Behauptung eines >literarischen Fingerabdrucks< geführt. Mit ihr ist freilich ein Schritt zu weit getan: Die Annahmen, daß der Sprachgebrauch eines Autors wie ein Fingerabdruck

übers ganze Leben hinweg unveränderlich bleibe, daß er sich ferner von dem eines jeden anderen zweifelsfrei unterscheide und daß er sich schließlich ohne größeren Aufwand nachweisen lasse - diese Implikationen der Analogie dürfen mit Fug und Recht als fahrlässig *ad acta* gelegt werden.

Es bleibt mithin festzuhalten, daß beobachtbare, meß- und beschreibbare Gewohnheiten in der Sprache eines Autors einerseits nicht in Frage gestellt werden können, daß jedoch andererseits die Extrapolation der hierüber gewonnenen Erkenntnisse, d. h. die Behauptung einer Konsistenz jener Gewohnheiten über das untersuchte Textkorpus hinaus, problematisch bleiben muß - eine Irrtumswahrscheinlichkeit, die sich einer genauen Bezifferung stets entziehen wird. Mit anderen Worten: Die literarische Kriminalistik wird sich damit abfinden müssen, vom Instrumentarium der Textquantifizierung nur Indizien, keine Beweise gewärtigen zu können.

Vor dem Hintergrund dieser Kautel umreiße ich im folgenden meine bisherige Beschäftigung mit der Frage, was und wieviel an Meß- und Beschreibbarem sich auf *einfache* Weise in Texten aufspüren und für die Identifizierung des Autors nutzen läßt. >E*infach*< sollte in diesem Zusammenhang dreierlei bedeuten:

- die ad-hoc-Verfügbarkeit der in Frage kommenden Verfahren, d. h. Verzicht auf vorbereitende Arbeiten wie die Erstellung eines Thesaurus etc.;
- Transparenz von Methode und Ergebnis;
- Reduktion des menschlichen Arbeitsaufwandes durch möglichst flexiblen EDV-Einsatz auf ein Minimum.

Das Ziel besteht nicht in der Lösung eines konkreten Identifizierungsproblems, sondern in der *Bereitstellung eines Instrumentariums* für solche Zwecke. Um dessen Verläßlichkeit während der Experimentalphase stets kontrollieren zu können, werden >Problemfälle< durch Umkehrung erzeugt: Aus vergleichbaren (1) Texten *bekannter* Autoren werden willkürlich Untersuchungsgruppen zu je drei Texten gebildet, von denen stets zwei vom selben Autor stammen. Konstruierte Aufgabe der Untersuchung ist es dann zu prüfen, ob sich die Zusammengehörigkeit jener beiden verwandten Texte gegen den dritten indizieren läßt. In der Zukunft soll auf der Grundlage genügend großer Erhebungen versucht werden, aus den Erfahrungswerten für die Affinität oder Divergenz von Textproben ein Substrat zu gewinnen.

Das Gattungsspektrum der bisher analysierten Textgruppen - darunter deutsche, französische, englische und lateinische - reicht vom Märchen über Belletristik, Philosophie und politische Pamphletistik bis zur Geschichtsschreibung und Politikwissenschaft. Die Größe der Texte schwankt, doch darf sie innerhalb einer Vergleichsgruppe nicht erheblich streuen; eine Menge von ca. 20 DIN-A4-Typoskriptseiten sollte als untere

Grenze nicht unterschritten werden; bestimmte Analyseansätze verlangen ein hoch wesentlich größeres Volumen. Aus den bislang erprobten, sehr zahlreichen Verfahren skizziere ich im folgenden nur einige wenige derer, die sich über *alle* Untersuchungsgruppen hinweg als aussagefähig erwiesen haben; die zur Illustration exemplarisch herangezogenen Zahlen und Graphiken sind dem Vergleich dreier kleiner französischer Texte entnommen. Als EDV-Werkzeug der Quantifizierung findet überwiegend TUSTEP Verwendung, dessen Systemcharakter hierfür von einzigartigem Wert ist: Kaum ein Frageansatz ließ sich *nicht* in kürzester Frist programmtechnisch realisieren, wo es mit anderen Mitteln - aufgrund der meist sehr schwierigen Formalisierung von Sprache - wochen- und monatelanger Entwicklungsarbeit bedurft hätte.

II.

Ein wichtiges Instrument im Dienste der gestellten Aufgabe ist nach wie vor der Wörterbuchvergleich. Als zweckmäßig erwies sich hierfür die spaltensynoptische Gegenüberstellung relativer Häufigkeiten zu jeder einzelnen Wortform in allen drei Texten, ggf. selektiv in vielfältiger Weise aufbereitet. (2) Es erübrigt sich, hier weiter darauf einzugehen, weil dieses Hilfsmittel zum einen schon hinreichend bekannt ist und sich in allen seinen Spielarten leicht realisieren läßt. Sein bedeutendstes Defizit jedoch ist, daß es intentional nur einen Steinbruch der menschlichen Interpretation verkörpert: Wie die Erfahrung zeigt, lassen sich darin regelmäßig für *jede* der möglichen Alternativen scheinbar schlagende Indikatoren finden. Zur Vermeidung solcher Aporien des interpretativen und subjektiven Zugriffs kann das Gesamtwörterbuch freilich auch der *reinen* Quantifizierung dienstbar gemacht werden: Für die drei Vergleichspaare (A-B, A-C, B-C) werden die Werte der relativen Häufigkeiten *aller* Einträge miteinander >verglichen< - will sagen: werden deren Differenzen ermittelt, die absoluten Unterschiedsbeträge summiert und die Endergebnisse als Kenngrößen ausgegeben [Abb. 1] Bereits bei der hier vorwaltenden geringen Textmenge (je ca. 20 Seiten) fällt ins Auge, daß die Summe der Differenzen zwischen den Texten B und C kleiner ist als die der anderen Vergleiche; im Regelfall verstärken sich die Unterschiede noch bei wachsender Textgröße. (Diese Beobachtung der Affinität von B und C bestätigt sich im folgenden immer wieder, weshalb ich künftig nicht mehr eigens darauf hinweise.)

Experimente mit der Filterung der Berechnung, d. h. Berücksichtigung nur der Einträge, die bestimmte Bedingungen (vgl. etwa Anm. 2) erfüllen, erbrachten kein höheres Signifikanzniveau. Im Gegenteil: Da solche Eingriffe das statistische Korpus schmälern, lassen sie lediglich die Irrtumswahrscheinlichkeit überproportional ansteigen.

Ein anderes Experiment erwies sich hingegen als erfolgversprechend: die Auswertung von n-Tupeln von Wörtern, d. h. von Einträgen, die nicht

Abb. 1: Vergleich der relativen Häufigkeit dreier Wörterbücher

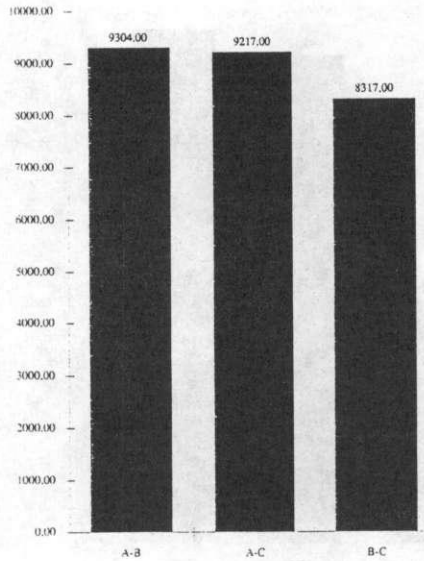


Abb. 2: Type-Token-Ratio (TTR), standartisierte Größe

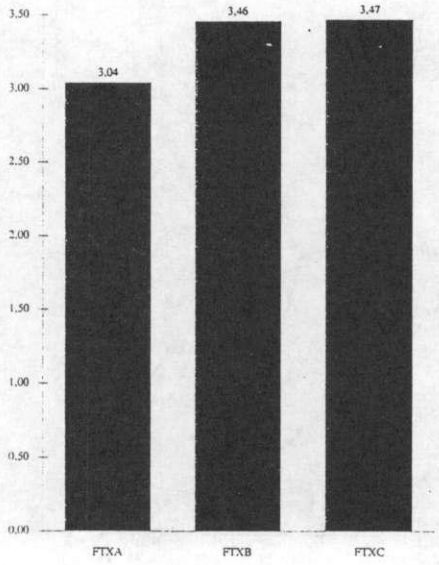


Abb. 3: Mittelwerte Silben/Satz

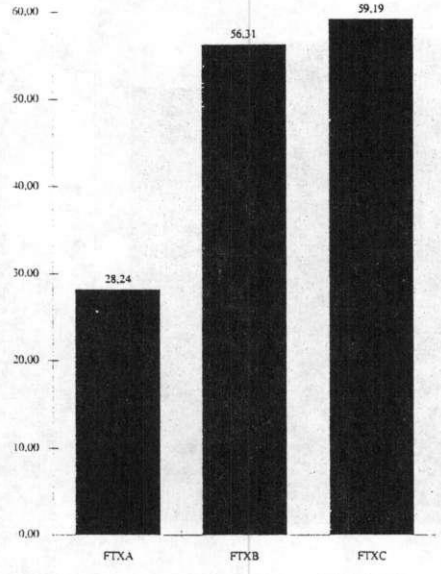


Abb. 4: Häufigkeitsprofil Wörter/Satz

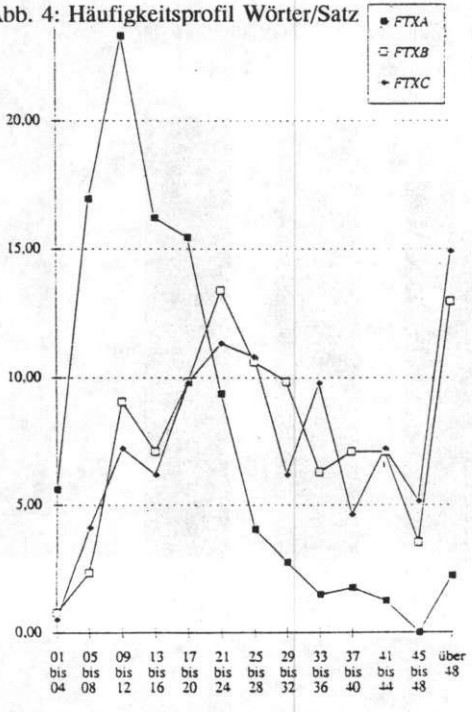
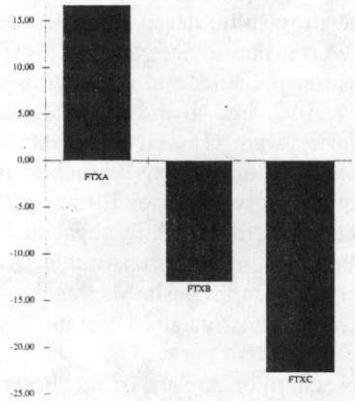


Abb. 5: Kontingenzkoeffizient Satzlängen (nach Buchstaben/Satz)



mehr von einzelnen Wörtern, sondern von *Wortverbindungen* auf Satzebene gebildet werden. Nachteil dieses Analyseverfahrens ist, daß erst die Auswertung sehr großer Textkorpora überzeugende Ergebnisse erbringt.

Ein weiteres probates Mittel ist die >Type-Token-Ratio< (TTR), bei der es sich i.w. um eine Kenngröße für die Breite des Wortschatzes in einem Text handelt. Unter den Möglichkeiten ihrer Berechnung sei hier die einfachste herausgegriffen: die Division der Zahl der in einem Text insgesamt vorkommenden Wörter (Tokens) durch die Zahl der darin enthaltenen *verschiedenen* Wörterbucheinträge (Types). Dieses simplifizierende Verfahren hat hier seine Berechtigung, denn da bei wachsender Textgröße der verwendete Wortschatz nicht linear mitwächst (Zipf's Law), wurde zur Vermeidung bestimmter Probleme die Berechnung auf der Grundlage standardisierter Textgröße (6332 Tokens in allen drei Texten) angestellt [Abb. 2].

Altbekannt ist auch die Ermittlung durchschnittlicher Satzlängen - meist nach Silben, seltener nach Wörtern oder Zeichen kalkuliert, obwohl es nach aller bisheriger Erfahrung für die Relationen nahezu bedeutungslos ist, welches Zählkriterium gewählt wird. Auch sie läßt sich maschinell leicht berechnen und erbringt oft überraschend deutliche Ergebnisse [Abb. *3*]. Indessen ist mit der Ermittlung solcher Durchschnittswerte ein Unsicherheitsfaktor verbunden: die jeder Mittelwertberechnung zugrundeliegende Nivellierung der Charakteristika eines Häufigkeitsverlaufs. Differenzierter und aussagefähiger sind deshalb Häufigkeit*sprofile* [Abb. 4], die maschinell nicht minder leicht zu errechnen sind. Daß sich analog dazu auch vieles andere auf Absatz-, Satz- und Wortebene quantifizieren läßt, was bemerkenswert eindeutige Hinweise auf Unterschiede und Übereinstimmungen von Textproben zutage fördern kann, sei hier nur angedeutet.

Im selben Zusammenhang habe ich auch verschiedene Strategien des selektiven Zugriffs auf die Häufigkeiten von Worttypen auf Satzebene erprobt. Alleinige Voraussetzung sollte deren formale Bestimmbarkeit über Präfixe, Infixe, Suffixe, Wortzusammensetzungen, Flexionsendungen oder auch bloß über frei definierte Laute und Lautkombinationen sein. Die Ergebnisse fielen dabei freilich um so indifferenter aus, je geringer der Anteil des jeweiligen Worttyps am Gesamtwortschatz der drei Texte war. Im ganzen gelang es - wie bei ähnlichen Versuchen im Wörterbuchvergleich - nicht, ein ubiquitär verwendbares Filter zu finden, das bessere Indizien als die schon demonstrierten hätte abgeben können. Auch hier gilt überdies, daß die Verringerung des statistischen Korpus die Irrtumswahrscheinlichkeit unverhältnismäßig erhöht, was die Aussagekraft eines scheinbar schlagenden Ergebnisses bis zur Unbrauchbarkeit diffundieren läßt.

Ein weiterer vielversprechender Ansatz ist die Berechnung eines Kontingenzkoeffizienten für die Abhängigkeit von Satzlängen untereinander. >Kontingenz< soll hier im statistischen Sinne ein Maß sein für jede Abweichung des Auftretens der Satzlängen von derjenigen Abfolge, die unter der Annahme stochastischer Unabhängigkeit zu erwarten wäre. Der Text wird zu diesem Zweck als Abfolge von >kurzen< und >langen< Sätzen untersucht (wobei der Median aller Satzlängen das Kriterium der Unterscheidung von >kurz< und >lang< abgibt) und geprüft, ob das Auftreten der Satzlängen einer Gesetzmäßigkeit unterliegt: Wäre dies nicht der Fall, müßten gleichnamige Satzpaare (kk & ll) annähernd ebensooft auftreten wie ungleichnamige (kl & lk). Der als Quotient zwischen der Differenz und der Summe der gleichnamigen und ungleichnamigen Matrixelemente errechnete Kontingenzkoeffizient müßte demzufolge im Bereich ±0% liegen. Ältere Untersuchungen von W. Fucks ergaben jedoch für *alle* von ihm analysierten Texte positive Werte, womit die These von einer Anziehungskraft von Satzlängen untereinander gestützt wäre. In der Tat trifft dies bei der hier vorgestellten Vergleichsgruppe französischer Texte auch für Probe A zu, überraschenderweise jedoch nicht für B und C: Deren Werte liegen beide deutlich im negativen Bereich [Abb. 5]!

Zum Abschluß dieser Skizze verdient noch ein selbst entwickeltes Turbo-Pascal-Programm erwähnt zu werden, das Texte auf der Grundlage eines Wortschatzvergleichs in den Einzelsätzen in Relation zueinander setzt. Weil dabei jeder Satz in der Quelldatei nach vorzugebenden Optionen mit jedem Satz der Vergleichsdatei verglichen werden muß, stellt dieses Programm höchste Anforderungen an die Leistungsfähigkeit der Hardware; außerdem ist leicht ersichtlich, daß nur kleine Textmengen (bis ca. 30/40 Seiten je Textprobe) in noch vertretbaren Zeiträumen bewältigt werden können. Der Aufwand lohnt gleichwohl, weil das Verfahren bemerkenswerte Ähnlichkeiten in den Formulierungen zweier Texte aufzudecken

vermag. Das Vergleichsergebnis - eine Dateiausgabe jener >Parallelstellen< - läßt sich seinerseits nach mehreren Kriterien quantifizieren, was nach allen bisherigen Erfahrungen ebenfalls vorzügliche Hinweise auf die Zusammengehörigkeit verwandter Texte ergibt.

III.

Auf den Versuch einer Generalisierung sei hier trotz der Stromlinienform mancher Ergebnisse verzichtet, weil das Volumen meiner *per diletto* geführten Untersuchungen noch zu gering ist. So wie im Exempel der drei französischen Textproben die vorgestellten bzw. angedeuteten Verfahren übereinstimmende Hinweise auf die Affinität der tatsächlich miteinander verwandten Texte erbrachten, (3) bewährten sie sich regelmäßig auch in anderen Fällen. Ohne jede Frage wäre es verfehlt zu glauben, daß in genügend vielen Messungen ein sicheres Maß zur Scheidung von Indiz und Beweis (oder auch nur zur genauen Bezifferung des Wertes solcher Indizien) gefunden werden kann: Letzte Instanz wird stets die philologische Kompetenz bleiben müssen. Andererseits deutet alles darauf hin, daß aus den unübersehbar vielen Quantifizierungsansätzen, die sich mit Hilfe des Rechners realisieren lassen, manch brauchbare neue Entscheidungshilfe für den mit Problemen der literarischen Kriminalistik befaßten Wissenschaftler zu gewinnen ist.

Anmerkungen

(1) Als Grundforderungen komparatistischen Arbeitens galt es einzuhalten, daß die *membra comparationis* derselben Sprache und Sprachepoche angehören und sich weder in der Literaturgattung noch im Sujet wesentlich unterscheiden sollten; überdies wurde der direkte Vergleich von Texten und Textfragmenten, die aus *einem* Werk genommen waren, ausgeschlossen.

(2) Beispiele solcher Selektivität sind: Übersichten über die in *allen* Texten oder über die in nur in *zweien* oder in *einem* Text belegten Einträge, über nur die mit einer bestimmten *Minimalhäufigkeit* vertretenen Einträge, ggf. nicht alphabetisch, sondern nach Häufigkeiten sortiert usw.

(3) Die Auszüge sind den folgenden Büchern entnommen: A: Maurois, André: Portrait de la France et des Français - 1955 [Kap. 1J B: Siegfried, André: De la III^e^ à la IV^e^ République - 1956 [Kap. 18]; C: Siegfried, André: Tableau des partis en France - 1930 [Kap. \\ Einen besonderen Hinweis verdient der Befund, daß die beiden Texte von Siegfried 26 Jahre auseinander liegen - eine bemerkenswert große Zeitspanne, in der der Autor seiner Sprache und seinem *modus scribendi* treu geblieben ist.