Erleichterung computergestützter Inhaltsanalysen durch verbesserte automatische Texterkennung

von Peter Hauptmanns und Rainer Schnell

Abstract

Ein Hauptproblem bei EDV-gestützter Inhaltsanalyse war bisher der Aufwand, der mit der Dateneingabe verbunden war. Neue Texterkennungssoftware zusammen mit preiswerten und leistungsfähigen Personal Computern erlauben nun automatische Texterfassung mit sehr kleinen Fehlerquoten auch für sozialwissenschaftliche Forschungsinstitute mit kleinem Etat.

Data input was one of the main problems of computer-assisted content analysis. New character recognition programs and low-priced powerful personal computers now permit high quality automated text recognition even for low budget research institutes.

Die aufwendige Dateneingabe für EDV-gestützte Inhaltsanalysen war bisher das technische Haupthindernis für einen breiteren Einsatz dieser Methode. Auch die in der Regel sehr kleinen Stichproben an untersuchten Texten sind eine Konsequenz des Eingabeaufwandes. Zumeist wird die Dateneingabe manuell durchgeführt, obwohl entsprechende maschinelle Techniken verfügbar sind. So gab ***Weber*** (1985:44) an, daß 100.000 Worte in 4 1/2 Stunden bei 0.1% Fehlern maschinell erfaßt werden könnten, ***Züll/Mohler/Geiss*** (1991:45) geben 50.000-200.000 Zeichen pro Stunde an. Allerdings sind die Preise für solche Geräte prohibitiv: ***Züll/Mohler/Geiss*** (1991:44) geben Preise von 30.000-40.000 DM an. Diese Kosten liegen aber außerhalb der finanziellen Möglichkeiten fast aller sozialwissenschaftlicher Forschungsinstitute.

Die technische Weiterentwicklung und der Preisverfall bei leistungsfähigen Personal Computern hat in Verbindung mit neuer Software diese Situation grundlegend geändert: Maschinelle Datenerfassung ist für Texte nunmehr mit weit geringeren Investitionen realisierbar. Möglich ist dies z.B. mit Programmen wie WORD-SCAN-PLUS, RECOGNITA oder OMNIPAGE. Um den praktischen Einsatz für die Dateneingabe bei Inhaltsanalysen beurteilen zu können, haben wir mit dem Programm WORDSCAN-PLUS der US-Firma CALERA einige Tests durchgeführt.

Neben dem Programm (ca. 700 US-$) und einem Flachbettscanner (ca. 2000 DM) wird nur ein dem derzeitigen Stand der Technik entsprechender PC (80486 CPU, 4 MB-RAM, 200 MB-Platte, VGA) samt Systemsoftware (DOS, Windows) benötigt: Die vollständige Ausrü-

stung liegt bei ca. 10.000 DM. Dies ist für fast jeden Institutsetat im Rahmen des Erträglichen, zumal die Ausrüstung darüberhinaus auch für die sonst üblichen PC-Anwendungen (Textverarbeitung, Datenanalyse, Datenverwaltung, Desk-Top-Publishing) einsetzbar ist.

Die Fehlerrate bei Texten von drucktechnisch guten Vorlagen lag bei den Versuchen im Mittel bei 1,7 Fehlern pro 100 Worte. Die meisten dieser Fehler (ca. 75%) ließen sich durch den Spelling-Checker des verwendeten Textverarbeitungsprogramms finden und beseitigen (WORDSCAN schreibt neben ASCII direkt auch eine Reihe von Textsystemformaten, u.a. MS-WORD und WORDPERFECT). Die Schrifttype der Textvorlage erwies sich dabei als weitgehend irrelevant für die Güte der Ergebnisse; eher von Bedeutung ist die Schriftgröße - bei Schriften kleiner als 8 Pkt. nimmt die Zahl der Fehler leicht zu. Das Programm muß nicht für jeden Buchstaben trainiert werden (wie beispielsweise das ältere SPOT von Flagstaff Engineering); z.B. konnte die englische Abstract-Seite der Kölner Zeitschrift für Soziologie und Sozialpsychologie mit nur zwei Fehlern pro Seite erfaßt werden. Die Fehler traten nur bei den kursiv gesetzten Namen der Autoren auf. Neben Zeitschriften, Zeitungsartikeln, Büchern, Faxausdrucken und Schreibmaschinentexten wurde auch ein Matrix-Drucker-Output (für diesen existiert ebenfalls ein spezieller Erfassungsmodus) gescannt und erfaßt[1]. In allen Fällen waren die Ergebnisse mehr als akzeptabel.

In einer "Testreihe" wurden acht Seiten aus verschiedenen, für Sozialwissenschaftler relevanten Quellen gescannt und vom Programm bearbeitet. Die Textvorlagen waren von unterschiedlicher drucktechnischer Qualität; weiterhin wiesen alle getesteten Seiten verschiedene Textformatierungsmerkmale (unterschiedliche Schriftgrößen, Fettschrift, Kursivschrift, Unterstreichungen o.ä.) auf. Bei den Vorlagen (1),(4) und (6) wurden mehrere (zwei bis drei) Versuche unternommen, um die optimale Scannereinstellung zu ermitteln (Helligkeit, Kontrast). Bei allen anderen Vorlagen wurde mit den Default-Einstellungen der Software gearbeitet. In der Regel erbringt ein längeres Experimentieren mit den Einstellungsmöglichkeiten Ergebnisverbesserungen.

Bei den Tests zeigten sich durchweg sehr gute Ergebnisse (siehe Tabelle 1); nur bei dem Versuch, eine Tageszeitungsseite mit schlechter Druckqualität zu erfassen, war die Fehlerrate relativ hoch. Allerdings wurden auch in diesem Fall 84% der Fehler durch den Spelling Checker gefunden. Nennenswerte Probleme gab es allein bei dem Versuch, eine Seite aus einem ZfS-Aufsatz von ***Diekmann*** (1990:269) einzulesen. Diese Seite enthält eine Reihe von Formeln mit griechischen Buchstaben, die vom Programm nicht korrekt erkannt wurden[2]. Außerdem wurden die Bruchstriche in den Formeln nicht als Textbestandteil interpre-

1 Nur der Vollständigkeit halber sei erwähnt, daß kein Programm existiert, das Handschriften erkennen kann. Das Programm eignet sich ebenfalls nicht zur Erfassung von Antworten in Fragebögen.

2 Allerdings waren diese Fehler systematisch: das "alpha" wurde immer als "a", das "lambda" immer als "k" erkannt.

tiert[3], so daß in diesem Fall Nachkorrekturen notwendig waren. Dieser Text wurde daher aus dem Test genommen. Texte dieser Art sind zur Zeit anscheinend nur mit größerem Aufwand maschinell erfaßbar.

Bei den übrigen Texten sind fast alle aufgetretenen Fehler auf Unsauberkeiten im Druck zurückzuführen[4]. Mehrspaltige Texte (wie z.B. bei der ZfS-Seite) werden vom Programm problemlos gelesen. Bei der Speicherung im ASCII-Format erfolgt die Trennung der Spalten über eine gleiche Anzahl von Leerzeichen. Bei der Speicherung im Format eines Textverarbeitungsprogrammes, daß zur Mehrspaltenverarbeitung fähig ist, erfolgt die Trennung der Spalten durch die entsprechende Formatierung.

Tabelle 1: Ergebnisse des Scan-Tests

Text	Worte	Fehler *	Fehleranteil	Bearbeitungs-zeit (sec.)
Zeitschrift für Soziologie	748	2 (2)	0,25 %	195
Kölner Zeitschrift für Soziologie und Sozialpsychologie	420	6 (2)	1,42 %	160
Buchvorlage (Sans Serif)	218	(-)	0 %	91
Frankfurter Rundschau	269	25 (21)	9,29 %	375
Fotokopie (Times Roman)	260	1 (1)	0,38 %	86
Offset-Druck (Times Roman)	258	7 (7)	2,71 %	244
ZUMA-Nachrichten	269	3 (2)	1,11 %	134
ZA-Information	308	3 (1)	0,97 %	99

* Durch den Spelling Checker (MS-WORD) erkannte Fehler

3 Senkrechte oder waagerechte Linien in der Vorlage werden bei der Texterfassung vom Programm als "Grafik" interpretiert und ignoriert.

4 Häufige Fehler waren z.B. die Erkennung des Buchstaben "m" als "rn" oder des Buchstaben "l" als "1".

Neben der Eingabe von Texten für die Inhaltsanalyse ist ein solches System vielseitig einsetzbar. So lassen sich ohne größere Mühe veröffentlichte Programme einlesen und dann kompilieren, wie Versuche mit abgedruckten FORTRAN-Programmen (die Vorlage bestand aus den Programmen von Lee 1980 und Loehlin 1987) und PASCAL-Programmen zeigten. Das Hauptproblem bei PASCAL-Programmen lag im Pointer-Symbol ("^"). Bei den drucktechnisch meist schlechteren FORTRAN-Programmen lag das Hauptproblem bei der Verwechslung von "J" mit ")" und "I" mit "J". Programmtexte lassen sich mit den heutigen Compilern rasch von solchen Fehlern säubern. Die Zahl der Fehler in Programmtexten lag mit 2-3 Fehlern pro Seite deutlich unter der Fehlerzahl, die durch programmierunkundige Datentypisten bei Programmen erzielt wird.

Ebenso von praktischem Interesse ist das Einlesen veröffentlichter Daten. Für rein numerisches Material existiert ein spezieller Erfassungsmodus, der hier für publizierte Korrelationsmatrizen und Datentabellen verwendet wurde. Eine Korrelationsmatrix (15*15) läßt sich nach dem Scannen in ca. 30 Sekunden in einen ASCII-File umwandeln, die Fehlerquote liegt bei ca. 1%. Dabei sind die meisten Fehler doppelte Vorzeichen oder Dezimalpunkte. Tabellendaten (ohne Dezimalzeichen, keine Vorzeichen) wurden nahezu fehlerfrei gelesen. Bei Daten mit Vorzeichen und Dezimalpunkt liegt die Fehlerquote ebenfalls bei ca. 1%. Der häufigste Fehler liegt in der Verwechslung eines Dezimalpunkts mit einem Dezimalkomma. In keinem einzigen Fall wurden zwei Zahlen miteinander verwechselt. Durch zweifaches Scannen desselben Materials können die Fehler rasch gefunden werden. In der Regel lassen sich mehr als 400 Dezimalzahlen pro Minute (ca. 1 Tabellenseite) so erfassen, mit Korrektur in ca. 2 Minuten.

Die Fehlerrate hängt sehr stark vom verwendeten Scanner und der Güte der Vorlage ab. So sind die Graphik-Fileformate "PCX" und "TIF", wie sie z.B. von Handscannern geschrieben werden, als Eingabe möglich. Allerdings konnten bei Versuchen mit Handscannern (eingesetzt: LOGITECH SCANMAN) nur für Zahlen bei jeder Vorlagengüte und für Texte auf außergewöhnlich gutem Material akzeptable Resultate erzielt werden. Mit einem Flachbett-Scanner (eingesetzt: HEWLETT PACKARD SCANJET+) lassen sich auch bei schlechten Vorlagen (Fotokopien, Faxe) gute Ergebnisse erzielen. Gegebenenfalls kann man durch einen modernen Kopierer vor dem Scannen die Qualität der Vorlage aufbessern. Die Geschwindigkeit des Einscannens lag bei ca. 20-30 Sekunden pro Seite, die Dauer der Texterkennung lag im Mittel bei ca. 2,5 Minuten pro Seite (auf einem PC 486-33). Dieser Wert variiert allerdings stark in Abhängigkeit von der Qualität der Vorlage (zwischen 1:26 Min. und 6:06 Min.). Da das Programm die Möglichkeit bietet, eine Liste von gescannten Files ohne interaktiven Eingriff abzuarbeiten, spricht nichts gegen das Scannen durch Hilfskräfte (so einfach wie Fotokopieren) und dem Erfassen als Batch. Die zusätzlich zur Hard- und Software entstehenden Personalkosten sind also sehr gering.

Literatur

Diekmann, A. (1990) :
Der Einfluß schulischer Bildung und die Auswirkungen der Bildungsexpansion auf das Heiratsverhalten. In: ZfS, Jg. 19, H. 4, S. 265-277.
Lee, E.T. (1980):
Statistical Methods for Survival Data Analysis. Belmont/California (Wadsworth).
Loehlin, J.C. (1987):
Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis. Hillsdale (Lawrence Erlbaum).
Weber, R.P. (1985):
Basic Content Analysis. Beverly Hills (Sage).
Züll, C./Mohler, P.Ph./Geis, A. (1991):
Computerunterstützte Inhaltsanalyse mit TEXTPACK PC. Stuttgart (Gustav Fischer)

Politischer Umbruch und Kriminalitätsentwicklung: Kongreß in Budapest

Die International Society for Criminology mit Sitz in Paris veranstaltet in Budapest vom 22. bis 27. August den 11. Internationalen Kriminologischen Kongreß. Kongreßsprachen sind Deutsch, Englisch, Französisch, Spanisch und Ungarisch. Der Kongreß wird organisatorisch in Plenarveranstaltungen, Vertiefungssitzungen und kleineren Arbeitsgruppen sowie mit ergänzenden Exkursionen durchgeführt werden. Das Generalthema lautet: Sozialer und politischer Umbruch und Kriminalitätsentwicklung - Eine Herausforderung auf dem Weg ins Dritte Jahrtausend -.

Die Plenarveranstaltungen widmen sich folgenden zentralen Unterthemen: (1) Grenzüberschreitende Strukturen der Wirtschaftskriminalität; (2) Umweltkriminalität; (3) Staats- bzw. Regierungskriminalatät und Korruption; (4) Terrorismus und Widerstandsbewegungen.

Für Vertiefungsveranstaltungen und Kleingruppen sind bisher rund 50 aktuelle Einzelthemen vorgeschlagen worden. Interessenten wenden sich bitte an:

NKG-Verbindungsstelle, Corrensstr. 34,
7400 Tübingen, Tel. 07071-292931.