Computer-assisted language analysis with the Macintosh

AMY HERSTEIN GERVASIO, JOHN TAYLOR, and STUART HIRSHFIELD
Hamilton College, Clinton, New York

Computer programs are useful in content analysis of written and spoken language. This article describes MacCALAS, a new and streamlined Macintosh version of CALAS, the Computer-Assisted Language Analysis System, originally designed in the early 1970s (Rush et al., 1974). MacCALAS, which is written in C, parses transcripts of spoken text into grammatical components and categorizes verb types according to Cook's (1979) case-grammar matrix. Using four subprograms, it determines the grammatical category of each word on the basis of adjacent function words and placement of words in a sentence, rather than by using a large "dictionary." MacCALAS also computes ratio measures of stylistic complexity and proportions of verb types. MacCALAS can be used in linguistic analysis of various kinds of texts and dyadic interaction such as psychotherapy and language development.

Content analysis of spoken and written text is frequently used to study differences in linguistic style. Computers can aid in quickly and accurately assigning linguistic categories. The Computer-Assisted Language Analysis System, or CALAS, designed in the early 1970s (Rush et al., 1974), is a mainframe system for parsing text into grammatical components and for categorizing verb types. It has been used to study the natural language of speakers in classrooms (Pepinsky, 1985), in psychotherapy (Meara, Shannon, & Pepinsky, 1979), and in assertiveness training (Gervasio, 1988). Although useful, the original CALAS programs are extremely cumbersome to use. This paper provides a brief chronology of the different computer programs used to analyze the text of conversation and describes MacCALAS, a new and streamlined Macintosh version of CALAS.

COMPUTERIZED ANALYSIS OF LANGUAGE IN PSYCHOTHERAPY

Beginning with the General Inquirer (Stone, Bales, Namenwirth, & Ogilvie, 1962), computer programs have been used to study the verbal content of psychotherapy. Programs by Harway and Iker (1964) and Starkweather and Decker (1964) were designed to compute the frequency of association of words in psychoanalytic sessions. These programs, which matched words in a text to a dictionary, required constant updating of the dictionary and were unable to handle polysemous words. Around this time Weizenbaum (1965) developed ELIZA, a program that made canned therapeutic remarks in response to human interaction.

John Taylor is now a graduate student in computer science at the University of Virginia, Charlottesville, VA. Requests for reprints should be sent to Amy Herstein Gervasio, Department of Psychology, Hamilton College, Clinton, NY 13323.

Winograd (1972) made an important innovation when he used rules of procedures rather than tables of words for the analysis of natural language. These programs were quickly adapted to studying the use of language in therapy. Clippinger (1977), expanding upon Winograd's methods, developed a variation of a computer program called ERMA that used rules of procedures to simulate psychoanalytic discourse. The Syntactic Language Computer Analysis System (or SLCA-III) of Cummings and Renshaw (1979) also incorporated Winograd's methods. The SLCA-III identified eight variables of "perceptual propositions of experience" that corresponded to parts of speech. Measures of the frequency of these categories have correlated with variables such as dogmatism and Machiavellianism.

CASE GRAMMAR AND MACCALAS

CALAS, developed at the Ohio State University by Rush et al. (1974), uses rules for analyzing language following a semantically based conception of "case grammar" (Fillmore, 1968). The focus of case grammar is on the inherent relationships between verb phrases and noun phrases and the roles that they play in conveying meaning in a sentence. As Bieber, Patton, and Fuhrman (1977) note, in psychological terms verbs say something about the various states, conditions, experiences, possessions, or actions of the things named and related to them. Measures of style, called stylistic complexity, often revolve around the ratio of verbs to other grammatical components. For example, average block length (ABL) measures the grammatical complexity of a text by dividing the number of clauses by the number of main clauses. Cook (1979) has shown that the ABLs of great writers correspond to the subjective impression of their stylistic complexity: Hemingway is less complex than Faulkner.

MacCALAS utilizes Cook's (1979) case-grammar matrix, in which verbs are divided into three basic types:

Table 1
Examples of Verb Types

Type	Basic	Experiencer	Benefactive
Stative	am	think (SEC)need (SEA)	own
Action	work	mention	send
Process	tire	see	found

Note—SEC = stative-experiencer-cognitive; SEA = stative-experiencer-affective.

stative, action, and process. Stative verbs describe a non-causal relationship between a person or thing (e.g., all variations of the verb "to be"). Action verbs describe a causal relationship with the specification of an agent who acts (e.g., do, run). Finally, process verbs describe a causal relationship in which something is happening to a person or thing without the specification of a causal agent (e.g., live, tire). These verb types can be "crossed" with the benefactive and experiential cases to produce a matrix, as shown in Table 1. (A case is the role that a noun takes in relation to a verb.) Verbs that take on the experiential case describe feelings, knowing, and sensing. In psychological research where the expression of feelings or cognitions may be deemed very important, stative-experiencer verbs are divided into stative-experiencer-affective (SEA) verbs such as "need or like" and stative-experiencer-cognitive (SEC) verbs such as "know or guess." Verbs in the benefactive case describe actions, states, or processes in which something is a beneficiary, such as "owe" or "receive." (For more information about the rationale for the original CALAS verb matrix, see Gervasio, 1984.)

MacCALAS PROCEDURES

In MacCALAS, as in CALAS, a transcript of spoken language is prepared in dialogue form and coded by speaker. MacCALAS uses one 800K disk, which contains five applications, two dictionaries, and two rule bases in one folder. It can be run on a Macintosh Classic. MacCALAS rules are written in C and modified from Snobol and PL/I, the languages used in the original CALAS. The first three applications in MacCALAS correspond to the original CALAS programs: Eyeball, Phraser, and Clauser/Caser.

Entering Text

There are two options for inputting text to be analyzed: (1) the text can be typed directly into a console window and parsed immediately, sentence by sentence, or (2) the text can be input using any Macintosh word-processing system, saved as "text," and chosen when the applications are used. The latter mode is preferred for long texts. (MacCALAS does have some specialized rules for entering questions, contractions, and phrases with embedded objects and adverbs.)

Four options for outputting text are available. The output can be (1) sent to the console, (2) saved as a text file, (3) sent to the console and saved as a text file, or (4) sent to the console and the printer. Thus, the user can see the results of the application as it is running. To correct the output, the user opens the saved text file using a word processor and makes any changes directly on the screen. A hard copy can also be printed for future use. The corrected file becomes the input for the next phase.

Eyeball

Eyeball assigns a grammatical class to each word in a sentence. The grammatical class is based upon "function" words: articles, auxiliary verbs such as "have," and prepositions. Because it uses position in a sentence rather than tables of words, the Eyeball dictionary needs only about 600 words. (The dictionary can be updated.) After grammatical categories are assigned, human editors familiar with rules of English grammar correct the output for each phase. Although Eyeball is accurate more than 80% of the time (cf. Gervasio, 1984), human editors are essential for handling polysemous words and idiomatic expressions. For example, the word "look" in the dialogue presented in Table 2 is not a verb that refers to seeing, but an "extra filler," much like the word "well." In MacCALAS, corrections to Eyeball and all other phases can now be made directly on the screen, rather than by running separate programs as was done in the original CALAS. (See Table 2 for an example of Eyeball phase with a dialogue with two speakers.)

Phraser

After Eyeball is corrected, Phraser groups words into phrases by using about 65 transformational rules (see Table 3). Phraser needs very little correcting by the human editor.

Clauser/Caser and Totaler

Clauser/Caser groups phrases into clauses. Main clauses are set off from subordinate clauses. Caser assigns a case role to each noun or verb in a clause. The human editor can change incorrect verb categories in this phase (see Table 4). The fourth application, called Totaler, is an addition to the original programs. As in CALAS, it keeps a running total of the number of sentences, clauses, phrases,

Table 2
Uncorrected Eyeball Phase of MacCALAS (Data Adapted from Gervasio, 1988)

*Speaker 1												
Dr Owens, look, I would like to talk to you about my work schedule.												
N	N		VE		N	X		V		T	F	P N P D NJ N
*Speaker 2												
Which reminds me, Helen, I am in kind of a bind.												
N		V		N		N		N	V	PA	I	A D N

Note—N = noun; X = auxiliary verb; V = verb; T = infinitive part of verb; P = preposition; D = determiner; J = adjective; A = adverb; E = extra/expletive; I = intensifier.

Table 3
Example of Phraser

*Speaker 1												
Dr Owens , look , I would like to talk to you about my work												
N		E		N	V			V		P		P
schedule .												
*Speaker 2												
Which reminds me , Helen , I am in kind of a bind .												
N		V		N		N		N	V	P		

Note—N = noun phrase; V = verb phrase; P = prepositional phrase; E = extra.

Table 4
Truncated Example of Caser Combined with Totaler

TEXT		CASE	TYPE	S	Cl	Ph	W

SPEAKER #1							
1 A	1 Dr Owens ,	*	N	1	1	1	2
	2 look,	*	E	1	1	2	3
	3 I	EXP	N	1	1	3	4
	4 would like	SEA	V	1	1	4	6
B 5	to talk	AE	V	1	2	5	8
	6 to you	OBJ	P	1	2	6	10
	7 about my work schedule .		P	1	2	7	14

Note—S = sentence; Cl = clause; Ph = phrase; W = word; EXP = experiencer case; SEA = stative-affective-experiencer verb; AE = action-experiencer verb; OBJ = objective case; N = noun; E = extra; V = verb; P = preposition.

Table 5
Truncated Example of Stylistic and Verb Measures

SPEAKER TOTALS PSYCHOLINGUISTIC MEASURES						
	Word per:					
SPKR #1:	Phrase	Turn	PMC	ABL	ACD	
	2.00	14.00	7.00	2.00	1.50	
SPKR #2:	1.57	11.00	7.00	2.00	1.50	
TOTAL TXT	1.79	6.25	7.00	2.00	1.50	
VERB CASES AND THEIR PROPORTIONS						
SPKR #1 Total	Action	Process	Stative	SEC	SEA	AE
2	1	0	1	0	1	1
	.5	0	.5			
SPKR #2 Total	Action	Process	Stative	SEC	SEA	AE
2	1	0	1	0	0	1
	.5	0	.5			
GRAND	.5	0	.5			
TOTAL 4	2	0	2	0	1	2
	.5	0	.5			

Note—PMC = phrases per main clause; ABL = average block length; ACD = average clause depth; AE = action-experiencer; P = process; S = basic stative; SEA = affective verbs; SEC = cognitive verbs. Note that not all available stylistic measures or verb types are listed in this table.

and words for each speaker and for the entire text. It also computes and displays the frequency and proportion of 11 verb types and computes eight ratio measures of stylistic complexity (Cook, 1979), such as average block length and words per turn. In the original CALAS, the frequency of verb types and stylistic measures were computed by hand (see Table 5).

The last application, Word Counter, is also new; it counts the frequency of any researcher-designated word strings in a text. Word Counter was developed to identify those idiomatic expressions that do not fit typical categories but may be useful to study.

Research Using MacCALAS

The accuracy of MacCALAS was compared to the original CALAS by using segments of Gervasio's (1988) original data on the language used in an assertiveness training film. MacCALAS reproduced the original analysis exactly, demonstrating its fidelity to the original CALAS. MacCALAS is currently being used in two projects: an analysis of the language used by well-known therapists in the ongoing film series *Three Approaches to Psychotherapy-III* (Shostrom, 1986) and an analysis of the language used in assertiveness training programs adapted for Native Americans.

Besides its use in research on therapeutic dialogue, MacCALAS could be helpful in other areas of psychology. In developmental psycholinguistics and in communications training programs, it could show the extent of change in such variables as length of utterance, content, and style. It could help identify presumed gender differences in stylistic complexity. It could also be used to illustrate stu-

dent use of particular grammatical and rhetorical styles in English composition classes and speech classes. In summary, MacCALAS can be utilized in the linguistic analysis of written text as well as the natural language of speakers in a variety of interactions.

REFERENCES

- BIEBER, M. R., PATTON, M. J., & FUHRMAN, A. J. (1977). Meta-language analysis of counselor and client verb usage in counseling. *Journal of Counseling Psychology*, 24, 264-271.
- CLIPPINGER, J. H. (1977). *Meaning and discourse: A computer model of psychoanalytic speech and cognition*. Baltimore: Johns Hopkins University Press.
- COOK, W. A. (1979). *Case grammar: Development of the matrix model (1970-1978)*. Washington, DC: Georgetown University Press.
- CUMMINGS, H. W., & RENSHAW, S. L. (1979). SCLA-III: A meta-theoretic approach to the study of language. *Human Communication Research*, 5, 291-300.
- FILLMORE, C. (1968). The case for case. In E. Bach & R. T. Harms (Eds.), *Universals in linguistic theory* (pp. 1-90). New York: Holt, Rinehart & Winston.
- GERVASIO, A. H. (1984). Computer-assisted analysis of conversation. *Behavior Research Methods, Instruments, & Computers*, 16, 158-161.
- GERVASIO, A. H. (1988). Linguistic analysis of an assertiveness training film. *Psychotherapy*, 25, 294-304.
- HARWAY, N. I., & IKER, H. D. (1964). Computer analysis of content in psychotherapy. *Psychological Reports*, 4, 720-722.
- MEARA, N. M., SHANNON, J. W., & PEPINSKY, H. B. (1979). Comparison of stylistic complexity of the language of counselor and client across three theoretical orientations. *Journal of Counseling Psychology*, 26, 181-189.
- PEPINSKY, H. B. (1985). Language and the production and interpretation of social interactions. In H. Fisher (Ed.), *Language and logic in personality and society* (pp. 93-129). New York: Columbia University Press.

- RUSH, J. E., PEPINSKY, H. B., MEARA, N. M., LANDRY, B. C., STRONG, S. M., VALLEY, J. A., & YOUNG, C. E. (1974). *A computer-assisted language analysis system* (Tech. Rep. No. OSU-CISRC-TR-73-9). Columbus, OH: Ohio State University, Computer and Information Science Research Center.
- SHOSTRUM, E. (Producer). (1986). *Three approaches to psychotherapy-III* [Film]. Santa Ana, CA: Psychological and Educational Films.
- STARKWEATHER, J. A., & DECKER, J. B. (1964). Computer analysis of interview content. *Psychological Reports*, 15, 875-882.
- STONE, P. J., BALES, R. F., NAMENWIRTH, A., & OGILVIE, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7, 484-497.
- WEIZENBAUM, A. (1965). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, 9, 36-45.
- WINOGRAD, T. (1972). *Understanding natural language*. New York: Academic Press.