DOCUMENT RESUME

ED 064 897 EM 009 969

AUTHOR Leton, Donald A.
TITLE Computer Simulation of Reading.
INSTITUTION Hawaii Univ., Honolulu.
PUB DATE Apr 72
NOTE 9p.; Paper presented at the American Educational Research Association Annual Convention (Chicago, Illinois, April 3-7, 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Computer Programs; *Reading Research; Reading Skills; *Simulation
IDENTIFIERS *Simuread

ABSTRACT

In recent years, coding and decoding have been claimed to be the processes for converting one language form to another. But there has been little effort to locate these processes in the human learner or to identify the nature of the internal codes. Computer simulation of reading is useful because the similarities in the human reception and perception of orthography and computer input allows such study. Computer simulation enables a more detailed study of the acquisition of reading skills than is possible in laboratory or classroom studies. In previous research a computer program was written to convert the word orthographies appearing in beginning readers to the segmental phonemes which define their oral representation. The computer program "Simuread" has now been extended to a third grade level of reading proficiency. The segmental phonemes are simulated by numerals, which are designated as phoneme equivalents. Program output illustrating the word processing is included here. (JK)

COMPUTER SIMULATION OF READING

Donald A. Leton
University of Hawaii

In previous research a computer program was written to convert the word orthographies appearing in beginning readers to the segmental phonemes which define their oral representation (Leton, 1969). The computer program, Simuread, has now been extended to a third grade level of reading proficiency. The segmental phonemes are simulated by numerals, which are designated as phoneme equivalents. Program output, illustrating the word processing, is presented in the handout.

Automated reading, to serve the educational needs of blind persons and to remediate the disabilities of reading handicapped, is within the range of possibility. There are a number of basic and technical problems, however, which first need to be resolved. An example of a technical problem is the need to improve the comprehensibility of artificial speech. An example of a basic problem in reading is to identify the graphic and oral language determinants of segmental and suprasegmental phonemes in the reading product.

In the past decade a teaching methodology, identified as the linguistic approach, has developed in the field of reading. This evolved from Bloomfield's (1942) recommendation that a child should be taught the alphabetic principle by teaching him the printed equivalents of his oral vocabulary. Gibson, et al. (1962) identified the grapheme-phoneme association as a learned unit in successful reading; and Chall (1965), Goodman (1968) and others, have referred to the process as coding.

Whatever claims may be made for the value of learning the alphabetic principle, it must also be recognized that the alphabet is the greatest error-producing device in the English language. Having 26 symbols, none of which consistently represent a single phoneme, each representing two to eight significant variations of sound, increases the risks of reading errors. The fact that each letter character has some phonemic value or values, and that most may conditionally have zero phoneme value also produces reading errors.

It would be inaccurate to imply that the alphabet behaves in a completely unprincipled manner in its relationships with oral language. Successful reading refutes such an implication. It is generally recognized that the conversion of the printed representation of the language to a corresponding oral representation is the essential definition of reading. Unfortunately, the views that the alphabetic principle, coding, or that grapheme-phoneme associations define this relationship are oversimplified.

The printed orthography includes punctuation marks, and logograms as well as the letters of the alphabet. Inadequacies of the alphabetic principle are illustrated in the following examples of orthography: <we're> is alphabetically identical to <were>, <does> is more commonly expressed as /dəz/ than as /dowz/, and <read> may convert to /r e d/ or to /r iy d/. The alphabetic principle has

little relevance to reading the following examples of orthography <@ 96¢ per doz. = 8¢ per oz.>, and <%, c/o, #, ≠>. The phoneme chains that are required for the numerals <1> and <2> are different from the phoneme chain <12>.

A grapheme, as it is traditionally defined is a minimum unit of graphics, i.e., each letter of the alphabet and each punctuation mark. Not all graphemes however convert to phonemes in oral reading. Some graphemes indirectly influence other grapheme-phoneme associations, e.g., the <e> in <rate>. The three graphemes in <the> convert to two segmental phonemes, and the three graphemes in <tax> require four segmental phonemes /t æ k s/ for its oral representation. The /ə/ phoneme in /ey b ə l/ is not designated by any of the graphemes in <able>.

From the field of psycholinguistics the terms coding, encoding, decoding, and recoding have been used to refer to the process of changing one language representation to another, e.g., aural to speech (Osgood, 1963) and visual to oral reading (Goodman, 1968). Chall (1965) recommends that coding be taught as a skill for beginning reading. There are erroneous assumptions as well as faulty definitions in these theories and teaching recommendations. There are at least three separate and interrelated codes in orthography: alphabet, punctuation, and logograms. There are at least two codes in phonemics: segmental phonemes and suprasegmental phonemes, i.e., stress, pitch and juncture. Some of the suprasegmental phonemes are determinable from the orthography whereas others represent oral patterning. When the term coding is used to refer to the relationship between the printed and oral representations of language it is not a single, nor a simple process, even for beginning reading.

The use of the terms encoding and decoding to explain reading and language processes may indicate an effort to utilize communication and information theories. Unfortunately, however, the terms are misused and sometimes defined incorrectly. Both Goodman (p. 17-19) and Osgood (p. 99-100) define decoding as a process of extracting meaning from the oral or the graphic representation of language, whereas, it is ordinarily defined as a procedure for converting a coded message into its prior language representation. Technically, it would be more correct to regard both the written and oral representations of language as hierarchical systems of codes, rather than as sources of information with meaning. The terms encoding and recoding used by these theorists refer to the coding of codes, i.e., of previous symbols and patterns of symbols. Devices such as telephone receivers and radio receivers decode patterns of electrical and radio waves. A listener may then comprehend, or extract meaning from the phonic codes of the oral language, but this is subsequent to the decoding.

Although coding and decoding are claimed to be the processes for converting one language form to another, there is little or no effort to locate these processes in the human learner, nor to identify the nature of the internal codes. This is a shortcoming in these theories. It would be necessary to identify the neuroanatomical structures, and the afferent-efferent and brain processes which produce the internal codes.

To refer to these briefly, there are visual and aural stimuli emanating from the graphic and phonic systems. These are transduced by the sense organs

and encoded in neuro-electrical impulses which are transmitted along the optic and cochlear nerve pathways, respectively. For the production of speech and script the efferent pathways transmit coded impulses to the muscles in the end organs for these expressions. Decoding, or the association and comprehension processes appear to be discrete and integrated cerebral processes, however. The perceptual, contral and motoral codes are not adequately objectified in the existing theories. The dysfunction of these processes may be clinically recognized in cases of dyslexia, dysphasia, and dysgraphia; but there is not a sufficient explanation of them for understanding normal language processing.

This critique of terminology and theories is not intended to discredit the teaching methods for developmental or remedial reading. Rather, a knowledge of theory limitations is regarded as necessary for a further understanding of the reading process.

The problem is reconceptualized as 1. the representation of orthography in interrelated graphic codes, 2. the segmentation or decomposition of orthography into graphic signals, 3. their association to phonic units, e.g., segmental phonemes, and 4. the modification of the phonemes by suprasegmental phonemes.

The computer simulation of reading is a double simulation problem. The orthography is simulated by holes in data cards and by patterns of electrical current in the computer. The holes in the data cards are keyed to represent the letters of the alphabet, punctuation, numerals and other logographs on a key punch. These holes are converted to electrical currents and are represented in on-off states, i.e., they are binary coded.

There are similarities in the human reception and perception of orthography and computer input. The human eye and the key punch can both be regarded as transducers. They transfer the orthography from the external environment and begin the process of internal coding. The energy in the light waves reflected from the orthography stimulate a photo-chemical process in the eye. The photo-chemical states are then converted into electrical states which are then transmitted on the optic nerve pathways to the occipital cortex. There are on-off states in the afferent pathways and in brain nerve cells on which the patterned electrical impulses are superimposed. Because of the coordinate nerve system, and the cerebellar structures and pathways which enable proprioception, the human reception of visual information is more complex than the computer reception.

The coding for the oral representation of language emanates in the motor cortex. The patterned electrical currents are transmitted on the efferent pathways and innervate the speech muscles. These operate the speech organs, to produce the oral expression of the language. The printer or output key punch and the speech organs are similar transducers. They transfer the energy from the electric current in the internal environment to the printer or the phonetic-acoustic code in the external environment. In this project the printer will simulate the oral language by printing predesignated numeral symbols for segmental phonemes. The human system for language expression is more complex, and more flexible than the computer system, but the computer holds a wide advantage in its rates for processing and output.

The ability to receive coded representations of orthography and the ability to produce language output are prerequisites to reading. They are characteristic abilities of the child-learner and the computer systems.

An entertaining question poses itself. Is it easier to teach a computer how to read, or is it easier to teach a child how to read? This leads to consideration of their internal abilities. They both present associative and memory capabilities to their teachers. The associative capability is necessary for relating the elements in the orthography codes to the coding apparatus for producing oral language. There is a marked difference in the nature of computer memory and human memory as these pertain to reading. In human memory there is a two-stage process of immediate recall and long-range memory. There are indications also that the visual, aural, tactual and motoral inputs into human's memory are processed in afferent and subcortical pathways so that certain features, or codes, are selectively enhanced or dampened, prior to their integration and inclusion in memeory. In contrast, the computer memory is a single state reservoir. Its potential use in artificial reading has not been evaluated as yet.

It may be equally easy, or equally difficult, to teach the computer and the child to read. The essential process is the same. The general procedures for the computer simulation which have been developed in previous research will be briefly described.

The minimum units of orthography, to which a segmental phoneme, a unitary phoneme-combination, or a non-phoneme is associated, are defined as signals. The phonemic associations are defined as designates. The alphabet characters, punctuation marks, numerals and other logograms are pre-specified as acceptable input. The computer program: 1) segments the orthography to identify its signal units, 2) tests for the existence of conditions in the word environment which determine its phoneme designate, and 3) outputs a pre-specified numeral equivalent of the phoneme designate. In Phase I of Simuread, i.e., learning and processing of primer words, the first occurrence of the signal did not require any identification of the environmental context for its initial association. As long as this sign-designate association continued in succeeding words the same phoneme output prevailed, and response-generalization form of learning was simulated.

When the signal required another phoneme designate, the conditions in the word environment which determined the second designate were identified. Thus, stimulus-discrimination learning was simulated in the program. The association to the second designate introduced response-discrimination learning.

Simuread is an algorithm for reading. The rules for each sign are developed as they occur in the texts of the readers. The algorithm is deterministic; the rules denote the sign-designates associations. The flow chart for Simuread is drawn on 1081 sheets of 8 1/2 x 11" paper, and the program contains about 4,000 Fortran statements. The program includes a procedure to compute the frequencies of associations and the conditional association matrix. This is a row matrix in which each row contains an independent graphic signal and each column contains a phoneme designate. Figure 1 in the handout indicates the procedures in Simuread.

A variation of the Trager and Smith phoneme system is used in this research. The phonic units include 33 segmental phonemes, 11 phoneme combinations, one juncture (suprasegmental phoneme) and a nonphoneme symbol for a silent grapheme, indicated on handout.

For the purposes of this presentation the following word corpora were processed: 1. Dale's list of 769 easy words for the Lorge formula of reading difficulty, 2. Houghton Mifflin's first grade readers and 3. Bank Street first grade readers. A statistical comparison of the grapheme phoneme associations in the Houghton Mifflin and Bank Street readers is presented in table 1 on the handout. The preponderance of similarity in the two matrices is due to the similar use of common English words; the differences in the frequencies of grapheme-phoneme associations stems from the differences in words and redundancies.

Authors and publishers of basic readers vary in their cognizance and control of word variables. They also show some awareness to the introduction of consonants, vowels and diphthongs in the vocabulary. Statistical analyses of the grapheme-phoneme associations which are required at various levels of a reading program have not previously been determined. These analyses provide basic information to show gradients in the acquisition of reading skills, i.e., the correct association of phonemes to the graphic units in words. This information may be of value for developing readers in which the grapheme-phoneme associations compare to a standard, for example, Standard American oral English. Or, the frequency may be controlled to reduce or increase certain phoneme associations for the reading instruction of bilingual or of language handicapped children.

Computer simulation of reading enables a more detailed study of the acquisition of reading skills than is possible in laboratory or classroom studies. The analyses of the conditional association matrix provides an objective means for identifying the learning-processing demands of various reading programs. There has been a tendency to grade beginning readers on the basis of word-vocabulary rate. Measures of reading difficulty are also based on word length and sentence length. The identification of the graphic signals, phonemic designates, their redundancies, and the environemental determinants of the associations are key variables in defining the acquisition of reading skills.

13.3

AERA
April, 1972

Computer Simulation of Reading

Donald A. Leton
Education Research and Development Center
University of Hawaii

Abstract

Numeric representations of phonemic units: The nine vowels /i e æ ɨ ə a u o ɔ/ are assigned numerals 1 through 9; twenty consonants /p t k b d g č ǰ f θ h v ð s š z ž l m n ŋ w r y/ are assigned numerals 10 through 33; /ks/ as one of the representations for <x> is assigned 34; eight vowel-semi vowel nuclei /iy ey ay oy uw ow aw/ are assigned 35 through 41; the combinations /yu/ and /wə/, as minimum representations for <u> as in: <use> and for <o> as in <one>, are assigned 42 and 43, the juncture phoneme /+/ is assigned 49, and the non phoneme symbol /Ø/ is assigned 50.

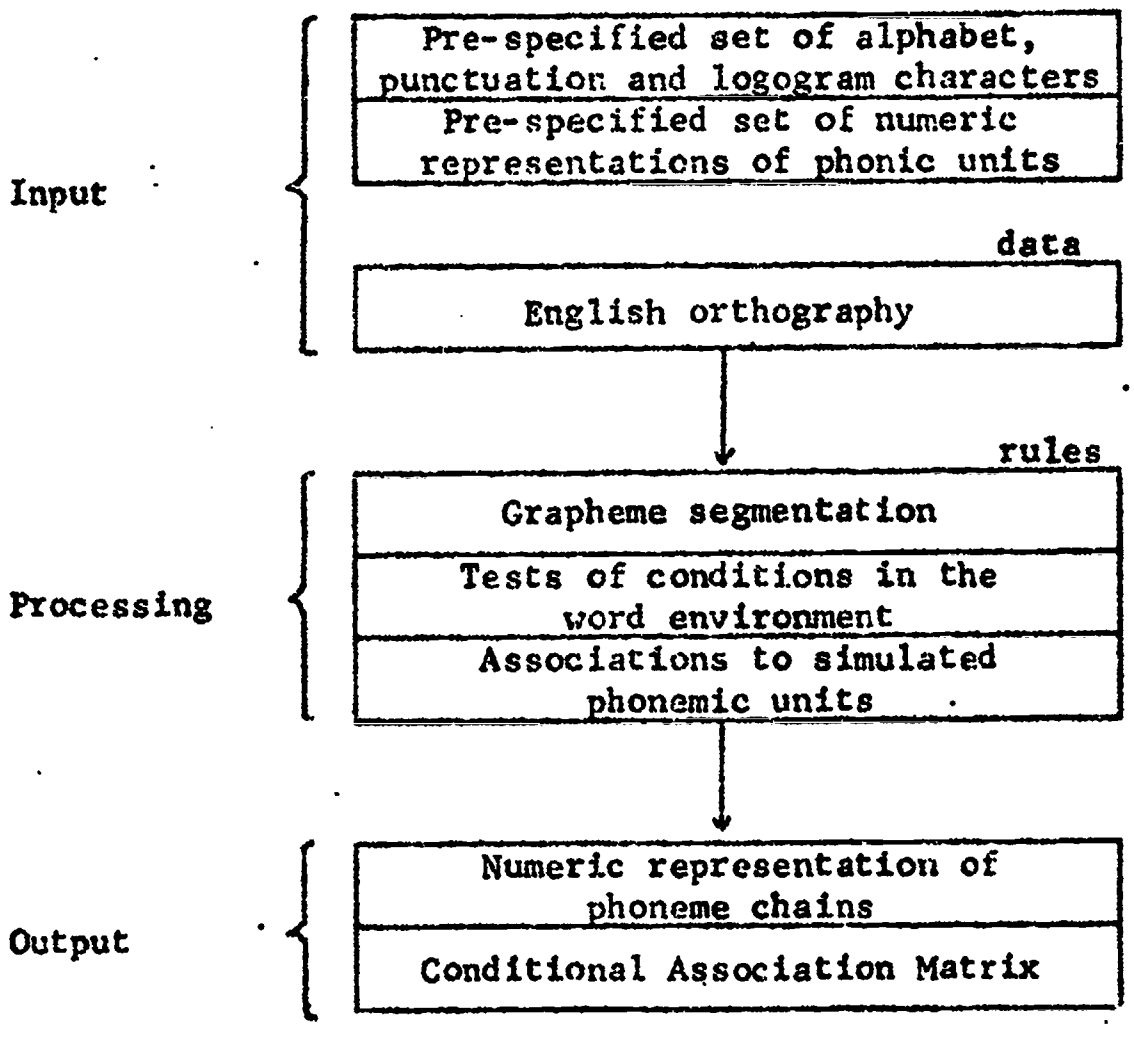

Figure 1. Chart of Simuread Procedures

ILLUSTRATION OF SIMUREAD OUTPUT

A

WORDS DO NOT SPEAK FOR THEMSELVES, WE ASK ONLY HOW TO GIVE *
4931 5321425491439492? 6114923103?124918 9324922 22823 22721 2234949313549 3231249 92927354920414911
394915 12150494949494949494949494949494949494949494949

THEM VOICE. SAY NOT THAT LANGUAGE AND LOGIC ARE BOUND IN MAN, *
4922 228492138235?4949233?4929 6114922 3114927 3303136175049 329144927 917 11249 63250491341291449 1
294928 32949494949494949494949494949494949494949494949

IMPART THEM TU A MACHINE TU LEARN THEIR LIMITS. THE TYRANNY *
49 12810 532114922 22849113949 54928 32435295049411394927 532294922 2324927 128 11123494922 549113732
32950354949494949494949494949494949494949494949494949

OF A MACHINE IS SEEN IN SLAVES WHO SERVE THEM. MAN IS BUT A *
49 52149 54928 32435295049 1254923352949 129492327 321 2254920394923 53221504922 228494928 32949 125
4913 51149 54949494949494949494949494949494949494949494949

SERVO DEVICE IN THE CULTURE OF TECHNOCRACY. *
4923 532214?4914 221372350?9 1294922 54912 52716 5325049 5214911 21629 91232362335494949494949494949
494949494949494949494949494949494949494949494949494949

B

BAND BANK BASKET BE BEAR BEAT BEAUTIFUL BECAUSE BED BEE BEEN BEFORE BEGAN *
4913 329144913 33012?913 32312 2114913354913 232491335114913?211 118 7274913 212 923504913 214491335
4913 1294913 218 5325?4913 215 32949494949494949

C

TIP CAN NOT HAVE YOUR BALL. YOU HAVE A BALL, TIP. GO HOME AND FIND YOUR BALL. *
4911 1104912 3294929 6114920 321504933 9324913 92750494933394920 3215049 54913 92750494911 110494915
404920402?5049 3291449183729144933 9324913 92750494949

D

CREAM," SAID ANNA. THE MAN GAVE ANNA THE ICE CREAM, AND *
491232352E4?454923 21449 32950 5494922 54928 329491536215049 32950 54922 549372350491232352?4949 329
144945454545454545494949494945494949494949494949494949

Sample A. Some philosophic statements, Sample B. Words from Lorge's list of 769 easy words,
Sample C. from Houghton Mifflin 1-1 Jack and Janet, Sample D. from Bank Street Reader 1st Grade
Uptown Downtoun.

Illustration of Rows in the Conditional Association Matrix

Rows extracted from the graphic-phonemic matrix indicating total occurences of graphic units and percentages of phoneme associations.

GRAPHS	FREQUENCIES OF PHONEME ASSOCIATES						NON-PHO	TOKENS
(B)	13							
(B)	0.0						1.0000	8
(C)	12	16	23					
(C)	0.7518	0.0	0.2082				0.0	538
(CC)	34							
(CC)	0.0							0
(CH)	12	16	24					
(CH)	0.1750	0.8250	0.0					120
(CI)	24							
(CI)	0.0							0
(CK)	12							
(CK)	1.0000							63
(D)	11	14	17					
(D)	0.0484	0.9520	0.0				0.0096	1672
(F)	18	21						
(F)	0.1277	0.7447					0.1277	47
(G)	15	17						
(G)	0.8354	0.1162					0.0484	413
(GH)	15	16						
(GH)	0.0	0.2462					0.7538	65
(A)	02	03	05	06	09	36		
(A)	0.0262	0.5731	0.1300	0.1300	0.0660	0.0746	0.0	1984
(AI)	02	03	36	05	37			
(AI)	0.5053	0.0	0.0947	0.0	0.0			338
(AY)	36	37						
(AY)	1.0000	0.0						139
(AU)	09	02						
(AU)	1.0000	0.0						17
(AW)	09							
(AW)	1.0000							48

Table 1

Row Totals of Graphic Units in Association Matrices for 1st Grade Readers

	H-M Jack & Janet	Up & Away	Bank St. Uptown-Downtown
Associations:	22323	29856	26759
Graphs		Row Totals	
<b>	90	13	8
<c>	249	619	538
<ch>	--	50	120
<ck>	286	176	63
<d>	1370	1886	1672
<f>	30	86	47
<g>	338	453	413
<gh>	10	36	65
<h>	578	899	717
<j>	380	178	116
<k>	481	390	472
<l>	1065	1459	1215
<le>	31	73	112
<m>	594	802	823
<n>	1397	1881	1818
<ng>	48	191	123
<p>	350	579	400
<r>	806	1412	1639
<s>	1272	1756	1812
<sh>	129	184	106
<t>	1942	2453	2000
<th>	664	913	1038
<v>	131	--	206
<w>	472	826	459
<wh>	183	147	96
<x>	29	33	46
<y>	540	609	612
<z>	--	33	3
<a>	1632	2367	1984
<ai>	338	392	338
<ay>	108	165	139
<au>	--	39	17
<aw>	48	60	48
<e>	2478	3208	2576
<ea>	37	103	89
<ee>	110	124	--
<ei>	--	--	19
<eo>	--	--	17
<ew>	19	11	27
<ey>	17	92	80
<i>	2028	1955	1549
<ie>	--	109	60
<o>	1230	1801	1429
<oa>	14	29	22
<oe>	19	19	8
<oi>	--	14	26
<oo>	195	284	344
<ou>	383	437	338
<ow>	--	24	502
<oy>	48	24	60
<u>	154	406	345
<ue>	--	56	--
<ui>	--	--	3

No tokens for graphs: <ci> <ph> <q> <rh> <sci> <si> <ti> <ye> <ae> <eau> <eu> <ia> <owa> <ua> <uy>