1965 International Conference on Computational Linguistics

SOME COMPONENTS OF A PROGRAM FOR DYNAMIC MODELLING OF HISTORICAL CHANGE IN LANGUAGE

Sheldon Klein

Carnegie Institute of Technology
Pittsburgh, Pennsylvania 15213
USA
and
System Development Corporation
Santa Monica, California
USA

SOME COMPONENTS OF A PROGRAM FOR DYNAMIC MODELLING OF HISTORICAL CHANGE IN LANGUAGE

Sheldon Klein

ABSTRACT

A system that is to serve as a vehicle for testing models of language change is being programmed in JOVIAL. Inherent in the design of the system is the requirement that each member of a speech community be represented by a generation grammar and a recognition grammar. The units of interaction in a simulation are conversations. Grammar rules may be borrowed or lost by individuals during the course of a simulation. The rules themselves need not be limited to those suggested by a particular theory of language; also, they may refer to any or all levels of linguistic phenomena. Extralinguistic factors pertinent to language change may be incorporated in simulations.

1.0 The Simulation System

A general simulation system which is to serve as a device for testing of hypotheses about language change through time is being programmed in JOVIAL, an ALGOL language, and is partially operational on the Philco computer (4).[1] The basic assumptions about the nature of language change inherent in the design of the program include the notion of generation grammar, Bloomfield's concept of speech community (1), and Sapir's concept of genetic drift (5). Aside from these built in concepts, the program is designed as a vehicle for testing models of language change as a function of variables selected at the discretion of an experimenter. It is intended that the simulation system be sufficiently flexible to work with either transformational or stratificational models of language; to simulate the interaction of members of a speech community among themselves and with members of other communities; to model special relations among particular members, e.g. family groups and social classes; to simulate multilanguage acquisition; and to model the transmission of language from generation to generation.

A basic assumption of the simulation system is that the interaction among members of a speech community is the prime

[1] This research is supported in part by Grant MH-07722, National Institutes of Health, United States Public Health Services (to the Carnegie Institute of Technology).

focal point of language change. Each member of a speech community or sample from a speech community is represented by both a generation grammar and a recognition grammar. Members of a community who are familiar with more than one language may be represented by additional grammars. The contents of the grammars may vary among individuals. Grammars of newborn children would be empty. An adult entering a new community as a speaker of an alien language might acquire an empty recognition and generation grammar to supplement the nonempty ones representing the languages he knows.

The basic units of interaction are speech forms produced in response to other speech forms. A primary function of the system is to simulate conversations among members of a speech community. During the course of a conversation, one individual will generate a form and another will attempt to parse it. Should the parser's rules be inadequate for the task, he may borrow the necessary rules from the generation grammar of the speaker, and perhaps use it when it is his turn to speak. Note that a bilingual speaker might attempt parsings with rules from all of his grammars.

Many decisions within the simulation system are made with the use of random numbers and functions governing the transition from one state of events to another. Monte Carlo techniques will be used in conducting simulations. Basically, the term refers to the use of random elements to solve essentially deterministic problems which may be too complicated to solve by deterministic methods. Accordingly, to evaluate the predictions of such a system, it is essential to determine the effects of different choices of random numbers numbers upon the results. If the model is deterministic, the results of repeated trials relying on different inputs of random numbers should be similar.

2.0 Components

The basic components of the the simulation system consist of a table containing the grammar rules and parameters associated with each individual in the simulation; a generation and parsing device that makes use of the grammars of interacting individuals; a table of functional relationships containing the rules of interaction pertinent to a particular simulation model; and, finally, a monitor program that determines the flow of the simulation and the passage of time, and that periodically takes a census to inform the experimenter of the changes occuring at various stages of the simulation.

The first version of the simulation system is being constructed around the author's automatic essay paraphrasing system (2) which produces essaylike paraphrases of an input consisting of a restricted English text and an outline of the desired output essay. The syntactic style of the output is controlled by manipulation of parameters pertaining to the frequency of usage of specific generation grammar rules (3).

The table of functional relationships that contains the definition of a particular model of language change might include rules expressing such features as:

1. Members of the same social group are more likely to speak to each other than to members of other groups.

2. Each time an individual interacts with a particular member of the community the probability of future interactions with that

member increases.
More complex functions pertaining to particular socio-cultural conditions might also be used.

Other functions might control the deletion of infrequently used grammar rules, or the shift of a grammar rule from a recognition grammar to a generation grammar.

The monitoring system is designed to work with a mixed assortment of functional relationships pertaining to very different phenomena. At a given decision point the monitor scans the table of functions sequentially until it finds an applicable item.

3.0 A Hand Simulation

The nature and function of the basic components can be illustrated by a hand simulation of the flow of an extremely simple language model.

Let the population contain six members: JOHN, MARY, HELEN, PETER, HERMAN and BABY. Let each have a separate generation and recognition grammar. Let each be assigned a status in the range of .01 to .99, and let the letters A,B,C,D,E,F represent the grammar rules existing in the community. (See table 1.) The content of the rules is deliberately left unspecified. The rules may refer to semantics, syntax, morphology and/or phonology. Each rule is associated with a weighted frequency. A rule with a frequency weight less than a specified threshold value (.1 in this simulation) can exist only in a recognition grammar. A rule with a frequency weight greater than or equal to the threshold must exist both in an individual's generation and recognition grammars. A rule existing in both grammars has the same frequency weight in each. A rule whose weight drops

	$T_{0,0}$	$T_{0,1}$	$T_{0,2}$	$T_{0,3}$	$T_{1,0}$	$T_{1,1}$	$T_{1,2}$
JOHN							
	S .8	S .8					
G							
	A .5	A .47					
	C .5	C .48					
	D .5	D .53					
R							
	A .5	A .47					
	B .04	B .02					
	C .5	C .48					
	D .5	D .53					
MARY							
	S .7	S .72	S .7	S .64			
G							
	A :5						
	B .5						
	D .5						
R							
	A .5						
	B .5						
	D .5						
	E .08						

Population
Table 1

	$T_{0,0}$	$T_{0,1}$	$T_{0,2}$	$T_{0,3}$	$T_{1,0}$	$T_{1,1}$	$T_{1,2}$
HELEN							
	S .4				S .4		
<u>G</u>							
	B .5				B .48		
	E .5				E .5		
					F .15		
<u>R</u>							
	B .5				B .48		
	C .02				E .5		
	E .5				F .15		
	F .06						
PETER							
	S .3				S .32	S .38	S .38
<u>G</u>							
	B .5						
	E .5						
	F .5						
<u>R</u>							
	B .5						
	D .08						
	E .5						
	F .5						

Table 1 Cont.

	$T_{0,0}$	$T_{0,1}$	$T_{0,2}$	$T_{0,3}$	$T_{1,0}$	$T_{1,1}$	$T_{1,2}$
HERMAN							
	S .6		S .6			S .6	
G							
	B .5		B .53			B .57	
	C .5		C .48			C .46	
R							
	B .5		B .53			B .57	
	C .5		C .48			C .46	
	D .02		A .07			A .05	
						F .05	
BABY							
	S .4			S .4			S .4
G							
							B .16
R							
				A .07			A .05
				B .07			B .16
				D .07			D .05
							E .05
							F .05

Table 1 Cont.

below a minimum value (.1 in this simulation) is deleted from all grammars.

Table 1 contains a record of the various states of the speech community at time $T_{i,j}$, where i refers to a major cycle--a single individual's interaction with a variety of speakers, and where j refers to a minor cycle--the interval of an interaction with a single speaker. At each increment in the value i, the monitor randomly selects a member as speaker for a major cycle. The monitor then scans the population sequentially to determine which members are to be auditors of the speaker. The determination follows the appropriate function contained in table 2. Each time an auditor is selected, the minor cycle time j is incremented by 1. When the monitor has scanned the entire community, the speaker's turn is over and a new one is selected to for the next major cycle. At the beginning of each major cycle the j or minor cycle value is set to zero. The data in column $T_{0,0}$ of table 1 are starting data supplied by the author. The data existing at $T_{i,j}$ is used in computing the state of events during $T_{i,j+1}$. Blank entries in table 1 indicate that the state of events is unchanged from the previous interval.

Table 2 contains the list of active rules refered to by the monitor during the course of the simulation. All computed values greater than or equal to 1 are rounded to .99; values computed at less than or equal to 0 are rounded to .01; in all cases, computed values are rounded to the second decimal place.

1. Probability of x speaking to y:

$$Ps_t(x,y) = \frac{.1}{/Status_{t-1}(x) - Status_{t-1}(y)/}$$

2. Frequency weight of recognition rule m at time t after use in parsing:

$$F_t(m) = F_{t-1}(m) - \frac{(F_{t-1}(m) - \text{relative frequency of m) in parsing at time t}}{5}$$

3. Frequency of rule not used in parsing at time t:

$$F_t(m) = F_{t-1}(m) - .02$$

4. Threshold frequency weight for adding or removing a rule from a generation grammar:

.1

5. Threshold frequency weight for removing rule from a recognition grammar:

.01

6. Status of speaker x after speaking to auditor y:

$$Status_t(x) = Status_{t-1}(x) - \frac{(Status_{t-1}(x) - Status_{t-1}(y))}{5}$$

Functions

Table 2

The simulation begins at time $T_{0,1}$ rather than at time $T_{0,0}$ for initialization purposes:

$T_{0,1}$

The monitor selects MARY as speaker for the 0 cycle, and examines the list of potential auditors. The first candidate is JOHN. According to function 1 of table 2 the probability of MARY speaking to JOHN is .1 divided by the absolute value of the status difference of the pair:

$$\frac{.1}{/.7 - .8/} \doteq .99 \qquad \text{(rounded)}$$

MARY will speak to JOHN because the random number generator of the monitor fails to yield a value greater than .99. Assume that MARY generates the form:

G(A,2D)

which is to be interpreted as indicating that, in the generation, rule A was used once, rule D twice. JOHN is able to parse the form with his own recognition rules, and their frequency weights are altered according to functions 2 and 3 in table 2. Rule A is computed as:

$$.5 - \frac{(.5 - .33)}{5} \doteq .47$$

Rule D as:

$$.5 - \frac{(.5 - .77)}{5} \doteq .53$$

JOHN's recognition rules B and C were not used in the parsing; after function 3 of table 2 each of their weights is decremented by .02. According to function 6 of table 2, MARY's new status becomes:

$$.7 - \frac{(.7 - .8)}{5} \doteq .72$$

$T_{0,2}$

The monitor searches for MARY's next auditor. MARY is skipped as a candidate. HELEN is next. The probability of MARY speaking to HELEN after function 1 of table 2 is:

$$\frac{.1}{/.72 - .4/} = \frac{1}{3.2}$$

Assume HELEN is rejected as an auditor because monitor's random number generator produces a value greater than this. Assume that the next auditor candidate, PETER, is also rejected. The monitor then selects HERMAN as the next candidate. Now assume that HERMAN is selected as auditor after appropriate computations. Let MARY's generated utterance be:

G(A,2B)

HERMAN must borrow rule A from MARY's generation grammar to complete the parsing. Rule A enters HERMAN's recognition grammar, by function 2 of table 2, with a value:

$$0 - \frac{(0 - .33)}{5} = .07$$

Since this value is less than .1, it does not enter HERMAN's generation grammar. The new value of B is computed as:

$$.5 - \frac{(.5 - .67)}{5} = .53$$

The rules not used in parsing are decremented by .02. HERMAN's recognition rule D, accordingly, drops below the minimum retention value of .01,and is deleted from his recognition grammar.

MARY's status is now computed as:

$$.72 - \frac{(.72 - .6)}{5} = .7$$

$T_{0,3}$

BABY is the next candidate for MARY's auditor. Assume that the monitor accepts BABY as a listener, and that MARY tells him:

G(A,B,D)

BABY must borrow every pertinent rule from MARY's grammar, each with a frequency weight,computed by function 2 of table 2,that is:

$$0 - \frac{(0 - .33)}{5} = .07$$

MARY's new status is now computed as:

$$.7 - \frac{(.7 - .4)}{5} = .64$$

The monitor has exhausted the list of candidates for auditor and a new speaker must be selected randomly.

$T_{1,0}$

Let PETER be selected as the new speaker. Assume that JOHN and MARY are rejected as auditors, but that HELEN is accepted:

G(E,F)

Rule E is in HELEN's recognition grammar and its new weight is:

$$.5 - \frac{(.5 - .5)}{5} = .5$$

remaining unchanged. The weight of rule F is computed as:

$$.06 - \frac{(.06 - .5)}{5} = .15$$

and after function 4 of table 2, F enters her generation grammar.

HELEN's unused rules are decremented by .02 .

PETER's new status is:

$$.3 - \frac{(.3 - .4)}{5} = .32$$

$T_{1,1}$

Assume HERMAN is picked as PETER's next auditor, and PETER says:

G(3B,F)

Rule B is in HERMAN's grammar and its new frequency weight is:

$$.53 - \frac{(.53 - .75)}{5} = .57$$

Rule F is borrowed from PETER's grammar and enters HERMAN's generation grammar with a value:

$$0 - \frac{(0 - .25)}{5} = .05$$

HERMAN's unused rules are each decremented by .02 . PETER's new status is:

$$.32 - \frac{(.32 - .6)}{5} = .38$$

$T_{1,2}$

Assume the monitor determines BABY to be the next auditor, and that PETER generates:

G(2B,E,F)

Rule B is in BABY's recognition grammar and its new weight is:

$$.07 - \frac{(.07 - .5)}{5} = .16$$

Accordingly, rule B enters BABY's generation grammar.

Rules E and F must be borrowed from PETER, and each enters

BABY's recognition grammar with a weight:

$$0 - \frac{(0 - .25)}{5} = .05$$

The rules not used in the parsing are each decremented by .02 . PETER's new status is:

$$.38 - \frac{(.38 - .4)}{5} = .38$$

4.0 Discussion

The preceding hand simulation should be sufficient to illustrate the operation of the simulation system. Anticipated computer simulations will involve 50 to 100 individuals, each associated with several hundred grammar rules. Unique parsings can be obtained by using existing frequency weights to determine preferential applicability of rules. The functions contained in table 2 can be greatly extended in number and content. One might wish to add special rules for interaction between parent and child, spouses, and among members of the same age group, etc., plus a mechanism for determining the birth and death of various members. The status factor might be divided into weights refering to social status, age, geographical proximity and the like.

The ideal test of the validity of a simulation is prediction. Hopefully, one might predict an attested state of a language from a model of an attested earlier stage. A major problem in such testing may be extreme sensitivity of a model to the choice of parameter values and constants. For example, the constants in the functions of table 2 seem to have the effect of making BABY learn too quickly. One might use a higher rate of decay for unused rules to decrease the learning rate. The need for trial and error manipulation of values will increase with the complexity of a model. Accordingly, one might start with simple models, increasing the complexity by stages.

The author's immediate research goal is to produce a stability simulation involving about 50 members,each associated with a simple phrase structure grammar of English, over a time span of 3 or 4 generations--a simulation in which the language at the start of the simulation is reasonably similar to the language existing at the conclusion.

References

1. Bloomfield, L. Language. New York: Holt, Rinehart, 1933.

2. Klein, S. Automatic Paraphrasing in Essay Format. In press, Mechanical Translation.

3. Klein, S. Control of Style with a Generative Grammar. In press, Language.

4. Klein, S. Dynamic Simulation of Historical Change in Language Using Monte Carlo Techniques. SP-1908, System Development Corporation, Santa Monica, December 1964.

5. Sapir, E. Language. New York: Harcourt, Brace, 1921.