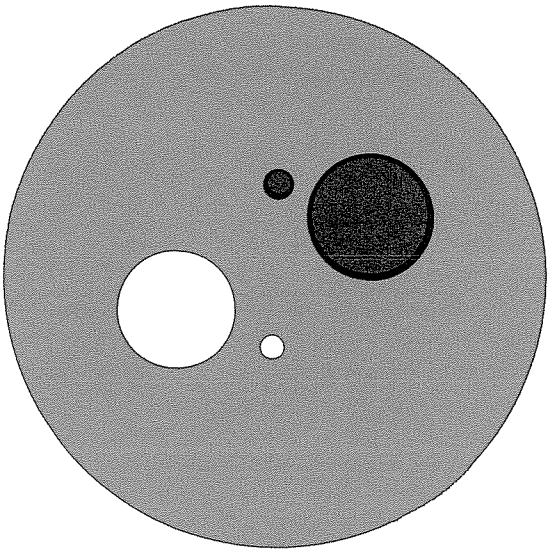
COMPUTER SCIENCES
DEPARTMENT

University of Wisconsin-
Madison



A COMPUTER MODEL FOR THE ONTOGENY OF PIDGIN & CREOLE LANGUAGES

by

Sheldon Klein & V. Rozencvejg

Computer Sciences Technical Report #238

December 1974

WIS-CS-238-74

COMPUTER SCIENCES DEPARTMENT
The University of Wisconsin
1210 West Dayton Street
Madison, Wisconsin 53706

Received: December, 1974

For presentation at 1975 International Conference on Pidgins & Creoles,
Hawaii, January 6-10

A COMPUTER MODEL FOR THE ONTOGENY OF PIDGIN & CREOLE LANGUAGES

by

Sheldon Klein & V. Rozencvejg

Technical Report #238

December 1974

ABSTRACT

A Computer Model for the Ontogeny of Pidgin & Creole Languages

Sheldon Klein & V. Rozencvejg
Computer Sciences Department
Linguistics Department
University of Wisconsin
Madison, Wisconsin 53706
USA

In this paper we describe both a computer simulation system for modelling language contact phenomena as a function of sociocultural, demographic and historical factors in general, and also a particular computer model for the generation and growth of Pidgin and Creole languages purely in terms of structural principles and mechanisms, without regard to the specific history of any particular Pidgin or Creole. (However, particular language situations may be used as test cases for validating our model.) The fact that it is possible to provide a structural-mechanism ontology model does not necessarily negate the validity of explanations that claim unique origin of Creoles. Rather, we seek to answer the question, "What would be the resultant Pidgin/Creole if language A were in contact with language B, given a communication language C (where C may be A, B or a third language) and where the native speakers of A are socially dominant-- and given the precondition that a Pidgin/Creole is to be the forced outcome of the simulation."

A generative semantic grammar is required for each language that is part of the model data. The computer simulation system can contain representations of individual speakers in the contact community (each associated with one or more grammars) interacting conversationally as a function of a predefined sociocultural-deomographic-historical model.

The process of interaction is viewed in our ontogeny model as a complex negotiation and bargaining phenomenon, where the desired end is communication, and where the tactics involve selection of and agreement upon constructions in the communication language that minimize the problems of semantic parsing for the participants.

The model is intended to be programmed in a meta-symbolic simulation system that includes:

a. a notational device for representating semantic deep structure in a network notation for generative semantic grammars of varying types.
b. a behavioral simulation language in which one may model rules of socio-cultural behavior for language communities. (A key feature is that the semantic deep structure of the non-verbal, behavioral rules may be represented in the same network notation as the semantics for natural language grammars, and, as a consequence, provide non-verbal context for linguistic rules.
c. a component for automatic inference of generative semantic grammars, and components for generation from and parsing to semantic deep structure representations, facilitating the modelling of language transmission across several generations.

A Computer Model for the Ontogeny of Pidgin & Creole Languages

Sheldon Klein & V. Rozencvejg *
Computer Sciences Dept.
Linguistics Dept.
1210 W. Dayton St.
University of Wisconsin
Madison, Wisconsin 53706

1. Background **

Portions of the automated model described in this paper are based on ideas and results of V. Rozencvejg as expressed in Rozencvejg (1969, 1972a, 1972b) and through private communications. The model itself is intended to be programmed in a meta-symbolic simulation system that includes:

a. a notational device for representing semantic deep structure in a network notation for generative semantic grammars of varying types.

b. a behavioral simulation language in which one may model rules of socio-cultural behavior for language communities. A key feature is that the semantic deep structure of the non-verbal, behavioral rules may be represented in the same network notation as the semantics for natural language grammars.

c. a learning component for automatic inference of grammars in the system.

Items a and b are already in existence in an integrated working system (Klein et al, 1972, 1973, 1974). A prototype of c exists as a program for learning transformational grammars (Klein & Kuppin, 1970). The learning component for generative semantic grammars is described in Klein (1973) and is under development. The reader is urged to read "Computer Simulation of Language Contact Models," (Klein, 1974) for an overview of the system; to read "Automatic Inference of Semantic Deep Structure Rules in Generative Semantic Grammars," (Klein, 1973) for details of the grammatical inference mechanisms; and to read "Modelling Propp and Lévi-Strauss in a Meta-symbolic Simulation System," (Klein et al, 1974) and "Automatic Novel Writing: a status report", (Klein et al, 1973) for details of the behavioral simulation system and the natural language generative component.

2. A Brief Survey of Components and Concepts

In the envisioned system, each individual in a modelled speech community may be associated with:

a. a semantic network of objects and relations describing his universe.

b. one or more generative semantic grammars that may accept portions of this semantic network as input to a surface structure production system.

c. a private set of behavioral simulation rules whose semantic deep structure is also encoded in the same semantic network (and in the same notation) as that associated with the natural language generative grammars.

Each indvidual also has a parsing capability which permits him to attempt to decode the semantic content of any sentences produced by other modelled speaker in the system. The decoding process may make use of any or all the grammars associated with the parsing individual and may refer to any of the information encoded in his semantic network.

The behavioral simulation rules govern the verbal and non-verbal behavior of each individual in the modelled community. The rules especially govern the patterns of verbal interaction among modelled speakers as a function of socio-cultural, economic and demographic factors. The rules themselves are formulated in an artificial language with its own grammar. Because the semantic

---

* V. Rozencvejg has not seen this draft of the paper, and the responsibility for possible misrepresentation of his views is the responsibility of S. Klein.

** Portions of this research have been sponsored by the National Science Foundation, The Wisconsin Alumni Research Foundation and The International Research and Exchanges Board.

deep structure of the rules is represented in the same semantic network notation as that for natural languages, it is possible to convert (in either direction) between verbal descriptions of behavior and non-verbal simulation rules governing that behavior.

There follows a brief description of terms used in this paper. The reader is again referred to the cited documentation for more detail.

semantic object An abstract entity, defined by its mappings into a lexical dictionary, its class memberships, and its functioning as an element of various semantic triples. Each semantic object in the system has a unique number. Objects may function as arbitrarily abstract or concrete entities.

semantic relation Similar to semantic objects, except that there is only one unique number for each type of relationship.

classes May include semantic objects or semantic relations as members. Class membership is dynamically modifiable in the system. A class may also be associated with and treated as a semantic object.

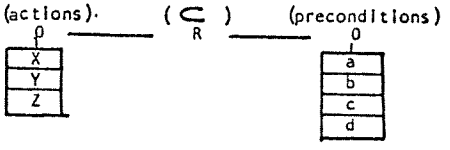
semantic triple Triples function as the basic units of which the semantic networks are composed. A triple may consist of a directed string of two or three objects and relations. Each triple in the network is assigned a unique number. Behavioral rules may create and delete triples. Each triple is associated with its time of creation and deletion.

predicate node Each semantic object in the network may also be linked to a list of numbers referring to certain other semantic triples in the network. The device permits a single semantic object to represent a complex idea or discourse. The device may be recursively self-referrential.

semantic deep structure of behavioral rules A rule consists of an action component and a precondition component. The actions may be the insertion or deletion of triples in the network. The preconditions may be tests for the existence or non-existence of specified triples. The preconditions may be viewed as the presuppositions of the the triples created by the action component. The action implementation may be a deterministic or weighted probabilistic function of the satisfaction of the

preconditions. The rules may be formulated in terms of specific triples composed of specifically named objects and relations, or they may be formulated in terms of classes of objects and relations (with greater generality).

The general form of the semantic deep structure of a rule is exemplified by the following: Let the preconditions for actions X,Y and Z be functions of the existence or non-existence of triples a,b,c and d. (In this rule, X,Y and Z are all consequences of the same precondition component.) Let the action portion of the rule be represented by a predicate node that is the 1st member of a triple. Let it be linked to another predicate node representing the preconditions by a relationship whose meaning is the logical connective "implied by" . The structure can be represented schematically as,



This type of notation makes the semantic content of the rules directly comparable and, in fact, part of the semantics for natural language grammar components in the same system.

surface structure//semantics rule

A surface structure//semantics rule consists of a phrase structure rule part plus a canonical form description of the type of semantic triple (or triple portion) that it may encode. Links from elements in the canonical form of the semantic portion to elements in the right half of the phrase structure portion indicate mappings of lexical items from the lexical dictionary that are also associated with the particular objects and relations united in a triple of the appropriate canonical form. The rules may be associated with frequency parameters, and they may be marked for usage in specific contexts defined by the simulation rules during the course of execution of the simulation model program. (The grammatical model also includes the use of various types of transformations.)

3. Goals of Parsing

Phrase structure syntactic parsing consists of reducing a string of terminal elements in a language to a single symbol by application of the rules in the reverse direction used in generation. The goal of semantic parsing is to decode the semantic content of a string or discourse without necessarily accomplishing a complete syntactic parse. In the system under discussion, a semantic parse may be considered complete if all the triples encoded in a sentence (or text) are determined. Within this constraint there are several modes of parsing possible.

Verification

The input string may be decoded with the criterion of verification that the string encodes triples that are already known to be true.

Plausibility

If the parsing yields triples that are not currently known to be true in the data structure, the parser in this mode can query the semantic deep structure of the simulation rules to determine if the triple structure implied by the string under parse might at some time be generated by the rule actions if the proper preconditions are met. Here acceptability of a parse includes testing of the satisfaction of the preconditions. The process is equivalent to automated presuppositional analysis.

Oracular Acceptance

The input string is taken as oracular and tutorial in this mode. Assuming an unambiguous interpretation is possible, the system would accept as true the triples implied by the string, and also accept as true any preconditions for those triples if rules for creating them also exist.

4. Types of Parsing

The simplest kind of semantic parsing can be accomplished, in principle, with no use of syntax whatsoever. The method is essentially a coordinate index information retrieval technique. Given a set of words, morphemes, etc., in a string,

one may locate the set of implied semantic triples while ignoring the ordering of the elements. All that is needed is an index containing back pointers to the semantic objects and relations that each lexical item could represent, and also an index indicating the semantic triples containg each particular semantic coject or relation. Given several unordered, implied semantic objects and relations, the set of all triples containing the items must include the ones actually encoded in the surface string under parse. Unfortunately, this method of semantic parsing is practical only for very short strings, for the implied combinations increases exponentially (or worse) as the length of the surface string increases. However, the method is rather practical for short sentences involving only two or three objects and relations. The technique is even useful as an intermediate step in a program for learning syntax when the mappings between the lexical items and semantic elements are known.

A mixed mode of parsing using some syntactic cues in combination with the coordinate index system is also possible. Here major syntactic markers of phrases in a long sentence can be used to partition the string into a series of groups of two or three objects and relations.

It is computationally implausible to envision a semantic parsing process that performs a complete multipath syntactic analysis of a string before attempting to decode the semantics. The only computational model that offers any possibility of modelling human speed in decoding must involve a mixed mode application of syntactic and semantic cues, where the primary tools are semantic criteria, and where detailed syntactic parsing is used where semantic heuristics fail or remain indeterminate.

5. Automatic Inference of Surface Structure//Semantics Rules

Detailed descriptions of the techniques are to be found in Klein & Kuppin (1970) (early model) and in Klein (1973).

In summary, the key learning heuristics involve frame matching between sentences, between sentences and rules, and between rules. Mismatches in the frames are candidates for combination in the same higher order class. Implied generalizations of rules are tested by generation of sentences derived from the newly posited rules. Acceptance or rejection of these sentences by other modelled members of the speech community controls the evolution of the grammar. (Acceptance implies that their grammars can parse the sentences.)

The learning model in Klein (1973) assumed that the semantic deep structure of sentences was known, even though the mappings between semantic objects and relations and their lexical representations was not given. We might also assume that the lexical items and their mappings onto semantic objects and relations are known, but not the semantic deep structure of sentences. That is, given the sentence, "dog bites man" one would know that there is a dog and a man and a biting, but not know who bites and who is bitten. The priming of our model must start somewhere, as we are not prepared to model language acquisition from phonology through morphology at this time. We assume undefined learning mechanisms have supplied some basic information. (For example, that some lexical items have been identified through assumed gesture specification, pointing, naming, etc. and that some meanings have been determined from the context of the social situation independent of the descriptive verbal input.) If the deep structure is given, together with morpheme boundaries, but with no specification of morpheme representation of semantic objects and relations, these may be inferred. (Actually the boundaries may enclose units larger than morphemes.) If the relations between lexical items or morphémes and the objects and relations they represent is given, but no information about their structuring into semantic triples is supplied, this information may be inferred. The learning component should be capable of working with either kind of data, or with any mixture of information for a given sentence.

A key learning heuristic is the coining of a rule or rules necessary to complete a parse where the embryonic grammar can supply only a partial parse. Where more than one partial parse of a sentence is possible, we posit a decision

criterion for selecting that incomplete parse that maximally simplifies the problem of semantic parsing.

6. The Pidgin/Creole Ontogeny Model

We wish to describe a computer model for the generation and growth of Pidgin and Creole languages purely in terms of structural principles and mechanisms, without regard to the historical reality of a particular Pidgin or Creole situation. Of course the behavior simulation language permits the incorporation of known non-linguistic, historical factors, and even the incorporation of posited but unattested factors. The possibility of deriving Pidgins and Creoles via a structural-mechanism ontology model does not necessarily negate the validity of explanations that claim unique origin of Creoles. Rather, we seek to answer the question, "What would be the resultant Pidgin/Creole if language A were in contact with language B, given a communication language C (where C may be A, B or a third language) and where the native speakers of A are socially dominant--and given the precondition that a Pidgin/Creole is to be the forced outcome of the simulation."

Consider first the case where initially there are only two languages involved. The model requires conversational interactions between members of the socially dominant members of group A and members of group B. The social dominace is reflected in the kinds and extent of language learning that will take place in the grammars of the members of the different groups. Complete application of all possible learning heuristics is ruled out for an initial language contact situation, and would only take place in child language learning. The exact profile of learning heuristics that would apply is a parameter subject to manipulation in computer experiments. The fact that speakers are adults and are learning through the framework of complete grammars of their native languages would automatically favor learning of constructions that harmonize with preexisting rules. The experimenter may have control of a number of factors. He might make the model require that members of group B accept lexical items from members of group A rather freely, and that members of group A be reluctant to accept lexical material

unless forced as a last resort heuristic in learning.

A major factor in the kind of grammar that develops is the training sequence. Because different training sequences can yield different forms of grammars, two different orderings of the same set of input sentences can yield two different grammars, each of which accounts perfectly for the sentences in that sequence, but each of which allows or implies different possibilities for sentences not yet encountered, because of the generalization tactics of the learning heuristics. A sociolinguistic behavioral model that determines who speaks to whom and who learns from whom can, accordingly, control dialect variation as a function of sociological parameters.

Another key control is selective ability to ignore inputs for reasons of excessive complexity. One could make the model deliberately ignore sentences of excessive length (as in child language learning where acceptable lengths could be a function of age). However the control could emerge automatically from a more basic principle: that sentences which do not yield to satisfactory semantic parsing are ignored. This more general principle would apply to adults as well as children. Such a requirement would force the initial acquisition of grammars favoring exceedingly simple constructions and only gradually permit the development of complex syntax, and even then complex constructions that are easily segmented into well defined simplex units would still be favored.

We note again that the hypothesis of this paper is that the dominant factor in language contact situations (and in particular Pidgin and Creole situations) is the learning of constructions that minimize the problems of semantic parsing for the participants. And we note that in our system semantic parsing includes decoding with reference to the semantic deep structure of the non-verbal behavioral simulation rules as well as that of the natural languages involved. Given our earlier discussion of parsing, it would seem obvious that short sentences are easier to parse than long ones, and that even in the absence of any syntactic rules whatsoever, our model would permit decoding of short sentences using the coordinate

index technique, and that once a semantic decoding is obtained the information can serve as inputs to automatic heuristics for determining relations to surface syntax.

The essence of the automated model is that it offers the potential for handling the language contact situation as a mass phenomenon--something impossible to analyze by hand calculation. Each training sequence would involve binary interac tion between specified members of each group, with controls on the grammars created as follows:

1. The A partner must be able to obtain a semantic parse of the test productions of his B partner. (Where either has dificulties, paraphrases may be requested and generated.
2. Rules learned by a member of B in an interaction with a particular A member must also be parsable by other A's.
3. Rules learned by an A in an interaction with a B must also be applicable in interactions with other B's.
4. Rules learned by a B must be consistent with rules learned by other B's and useful for inter-B communication.

A situation involving three languages (where the initial contact language is not the first language of either A or B) is not essentially different. Precisely the same mechanisms are at work. The system permits learning of context applicability of rules, as well as permitting maintenance of separate grammars by an individual If a modelled speaker is treated as having two separate grammars, he may learn rules in one using the structure of the rules in the other as a controlling factor, yet maintain the integrity of that grammar. In this system, both or even mixed modes are possible (in one sense, separate grammars and single grammars with social-context marked rules would appear to be logically equivalent devices). It seems likely that the two language simulation model and the three language model would eventually converge as a third communication language emerged from the two language situation.

The modelling of Creole ontogeny is essentially dependent on introducing newborn children into the speech community. Having no predetermined grammars to build upon, the grammar of children is entirely dependent on training sequence which may include sentences in the original language of B, an emergent A-B Pidgin and some use of the first language of A. The learning heuristics available to children are, as mentioned, much more powerful than those for adults in the model. The language transmission process may be permitted to continue for several generations, at least until none of the original A and B members of the first contact situation are left alive.

7. Discussion

What we have supplied here is a description of an automated meta-linguistic testing device that would permit the formulation and testing of language contact models in conjunction with sociocultural and demographic factors. The model is familiar to Creolists. What we have supplied is a formulation of these ideas in a notation that makes them amenable to empirical test. In a sense, the simulation system permits a specification of mechanism at a lower level than currently prevalent. We do not claim that the exact model and choice of heuristics suggested in this paper are the correct ones; rather that they are an example of possible sets of choices that may be subject to verification in automated testing. We do predict that for a given semantic parsing program (with a particular set of heuristics and tactics) there will be a specifiable definition of simplicity; and that the same set of languages should yield essentially similar Pidgins and Creoles in variant experiments that preserve the basic sociological matrix of Pidgin/Creole ontogeny.

The question of the existence of a universal semantic component remains. Our system is prepared to handle models involving both a universal semantic component and ones involving separate semantic structures for different languages.

We offer the hypothesis that the principle of ease of semantic parsing in language acquisition models might, in itself, account for the empirical evidence for the existence of a universal semantic component. If the hypothesis is correct, it would mean that a universal semantic component is not an innate genetic artifact, but rather a logical-mathematical artifact of the process of language acquisition.

8. References

Klein, S. 1973. Automatic inference of semantic deep structure rules in generative semantic grammars. Univ. of Wisconsin Comp. Sci. Tech Report 180. Also in press, Proceedings of 1973 International Conference on Computational Linguistics, Pisa.

Klein, S. 1974. Computer simulation of language contact models. In Towards Tomorrow's Linguistics, Shuy & Bailey, editors, Washington D.C.: Georgetown University Press.

Klein, S., Aeschlimann, J.F., Balsiger, D.F., Converse, S.L., Court, C., Foster, M., Lao, R., Oakley, J.D. & Smith, J. 1973. AUTOMATIC NOVEL WRITING: a status report. Univ. of Wisconsin Comp. Sci. Tech Report 186. In press (abridged) Proc. Int. Conf. on Computers in the Humanities, 1973, Minneapolis.

Klein, S., Aeschlimann, J.F., Appelbaum, M.A., Balsiger, D.F., Curtis, E.J., Foster, M., Kalish, S.D., Kamin, S.J., Lee, Y., Price, L.A., & Salsieder, D.F., 1974. Modelling Propp and Lévi-Strauss in a meta-symbolic simulation system. Univ. of Wisconsin Comp. Sci. Tech Report 226. In press, Patterns in Oral Literature, Jason & Segal, editors, 1973 World Conference of Anthropological and Ethnological Sciences, Chicago, Series. The Hague: Mouton.

Klein, S. & Kuppin, M.A. 1970. An interactive, heuristic program for learning transformational grammars. Computer Studies in the Humanities and Verbal Behavior. 3:144-162.

Klein, S., Oakley, J.D., Suurballe, D.A. & Ziesemer, R.A. 1972. A program for generating reports on the status & history of stochastically modifiable semantic models of arbitrary universes. Statistical Methods in Linguistics 8:64-92

Rozencvejg, V. 1969. Инфинитивные конструкции и балканские языковые контакты. Slavia, Rocnik 38, Sesit 2 (Praha).

Rozencvejg, V. 1972a. Основные вопросы теории языковых контактов. Новое в лингвистике: Выпуск VI: Языковы контакты. В Розенцвейг, редактор. Москва: издательство "Прогресс".

Rozencvejg, V. 1972b Языковые контакты. Академия наук СССР. Ленинград: издательство "Наука".