American Journal of Computational Linguistics *Microfiche 36 : 40*

Developing a Computer System to Handle Inherently Variable Linguistic Data

D. Beckles, L. Carrington, and G. Warner in collaboration with

C. Borely, H. Knight, P. Aquing, and J. Marquez

Department of Mathematics and School of Education
University of the West Indies
St. Augustine, Trinidad

ABSTRACT

Linguistic communication in Trinidad and Tobago is characterised by intra- and inter-ideolectal variation in a spectrum ranging from Creole-English to Internationally Acceptable English. The tape-recorded speech of a sample of children is being analysed to determine the structure of their language, its correlation with socio-linguistic factors and their progress in the use of English. The computer system is designed to deal with manually codified data in the form of parse trees with associated grammatical and semantic information. The communication complex does not have readily identifiable norms. The analytical method and computer system effect recognition of stable sub-systems (regardless of the external criteria which determine these sub-systems), comparison of these sub-systems with English as well as state the evolution of the children's language.

Acknowledgement

The research of which this paper is a working document is partially funded by Ford Foundation Grant 690-0664D. The authors acknowledge the kind assistance of the IBM World Trade Corporation, Port of Spain, Trinidad.

Preliminary

The design and some results of the research to which the computer system relates are described by Carrington, Borely and Knight (1969, 1972, 1974 a + b). Part of the intention of the project is to describe in terms applicable to curriculum development and teacher education, the structure of the speech of school-children aged 5-11+ in Trinidad and Tobago and to compare this speech with English.

The official language and medium of instruction is English. However, the medium of daily communication ranges from a type of Creole-English to a modifed variety of Internationally Acceptable English (IAE). The term "post-creole dialect continuum" has been used by several researchers, notably Le Page (1957), De Camp (1971) and Bickerton (1973) to refer to apparently analagous situations in Jamaica and Guyana. In addition to Creole, English and variants of both, a large part of the population is exposed to a local variety of Hindi (Bhojpuri). Smaller numbers are exposed to Lesser Antillean French Creole and fewer still to Spanish.

Communication within the society is characterised by inter-ideolectal variation related to several socio-linguistic factors - ethno-linguistic background, social class, educational level, occupation, sex and age. Code-switching and intra-ideolectal variation related to the context, content and purpose of communication complicate the examination of the communication system. Since the variant levels of the complex appear to overlap they are difficult to separate into distinct sub-systems.

The Linguistic Data

The available corpus comprises 100 hours of the recorded conversation of almost 1,000 children between 5 and 11+ selected randomly from 30 schools. The data fall into two pre-determined categories: (a) free (with peer group);

(b) controlled (with investigator). Given the nature of the communication complex stated above, variation and contrast are central to the data. In addition to the usual socio-linguistic correlates of variation, these data have the possibility of containing linguistic elements which are not paralleled anywhere else in the community. These elements may occur as a result of the instability intrinsic to the performance of a vulnerable age cohort. We are not dealing with fully learned discrete languages or dialects but with partially learned systems of speech communication being used by children who, by virtue of being in school, are under pressure to abandon part of their communication repertoire in favour of another variety of speech.

Implications of the Data Type for the Analytical Procedure

English is the only code of the communication complex for which adequate grammatical descriptions are available. It is demonstrably untenable to assume that the informants are attempting to speak English at all times. They are communicating in a set of language varieties which are assumed to be rule-governed. A statement of frequency and type of deviation from English cannot therefore be an adequate analysis. The first task of the analysis must be to determine the structures, both major and minor, used by informants of various socio-linguistic descriptions.

A preliminary examination of the data shows that at the level of phrase-structure of utterances, the structures will appear to be predominantly identical with English. It is the components of the elements, their meanings and functions that will show the differences from English. Consequently, the analysis must note the levels at which derivational trees cease to be compatible with English.

In view of the variability inherent in the data, the analysis must

discover the socio-linguistic correlates of the occurence of elements, as well as state co-occurence restrictions of a given element. Since it is possible that some elements may be distributed in a way that does not permit correlation with the stated socio-linguistic factors, the analysis must permit grouping of informants based on shared linguistic features for subsequent re-examination. This provision admits the possibility that sets of features may be typical of a language acquistion stage of the informants regardless of their socio-linguistic descriptions.

The Analytical Procedure

1. Each utterance is phonetically transcribed and ascribed to an informant by an identification procedure. Doubtful identity is specially coded.

2. Each utterance is rewritten in English orthography.

3. For each utterance a parse tree is constructed using the following protocol where each category described below forms the content of a node of the parse tree. The numbers are for reference and indicate the hierarchical relationship of the nodes.

Ø.Ø Utterance type	S	sentence
	SEL	elliptical S
	FRAG	fragment
	FREL	elliptical FRAG
Ø.1 Utterance complexity	SIMP	simple
	CP	compound
	CX	complex
	CPCX	compound-complex
Ø.2 Structural type	DEC	declarative
	INT	interrogative
	IMP	imperative

Ø.3 Semantic type

STMT	statement
QU	question
COMM	command
RHET	rhetorical intent

Ø.4 Linear order and type of clauses occurring
e.g. MC1 + ADVC TEMP 2

Ø.5 Linear order and type of phrases occurring
(where not part of a clause)
e.g. PREP P 1 + VBL P 2

Ø.6 Dependency of clauses - dependent
embedded
co-ordinate
included
e.g. 2/1 = clause 2 is embedded in clause 1

Ø.7

		ACTV	active
AFM	affirmative	PAS	passive
NEG	negative	EQ	equational
		STAT	stative
		LOC	locative

1.Ø surface structure of the clause/phrase occurring first.
e.g. MC1 → SUBJ + PRED* + IOBJ + DOBJ + PREP P
*PRED = predicator not predicate

1.1 detailed analysis of first occurring element of
1.Ø. e.g. SUBJ → PRMD + HDW

1.1.1 first element of subject. e.g. PRMD → [HE] PADJ,
RD, MASC, SG, NOK; IAE: [HIS] etc.

2.Ø surface structure of the clause/phrase occurring
second... etc to 7.Ø.

As exemplified at 1.1.1, the last node of each sub-part states the actual literal being described. The acceptability of the item as IAE is noted, OK or NOK, together with a reasonable IAE alternative. Apart from the obligatory information required by the procedure, the analyst may make additional comments which may be either in keywords or English. e.g. CMNT: probably idiosyncratic or CMNT: double NEG.

8.Ø is reserved for special idioms.

e.g. 8.Ø [SCRUNT] ⟶ scrounge for a living

9.Ø is reserved for tags.

e.g. 9.Ø TAG ⟶ [YOU HEAR]

Fig. 1 shows a sample analysis.

Figure 1

Ø652Ø72

[MY SISTER AND THEM DOES BREAK A SET OF PLATE, YES]

Ø.ØS; Ø.1 SIMP; Ø.2 DEC; Ø.3 STMT; Ø.4 MC + TAG; Ø.5 NA; Ø.6 NA; Ø.7 AFM ACTV

1.Ø MC ⟶ SUBJ + PRED + DOBJ

1.1 SUBJ ⟶ PRMD + HDW

1.1.1 PRMD ⟶ [MY] PADJ, ST, SG, OK

1.1.2 HDW ⟶ N. ASOC, ANIM, NOK: IAE: NEQV

1.1.2.1 N ASOC ⟶ NCO + ASOC

1.1.2.1.1 NCO ⟶ [SISTER] N SG, ANIM, OK

1.1.2.1.2 ASOC ⟶ [AND THEM] NOK; IAE: NEQV; VIDE 8.Ø

1.2 PRED ⟶ AUX + VT; GR @ CTN, @ PROG, PATT, NEUTTM

1.2.1 AUX ⟶ [DOES] NOK; IAE: ZERO

1.2.2 VT ⟶ [BREAK] OK TRAN

1.3 DOBJ ⟶ PRMD + HDW

1.3.1 PRMD ⟶ IND DET + N + PREP

1.3.1.1 IND DET ⟶ [A] OK

1.3.1.2 N ⟶ [SET] NCO, SG; LEX: NOK; IAE: [LOT]

1.3.1.3 PREP ⟶ [OF] OK

1.3.2 HDW ⟶ [PLATE] N PL, INAN, NOK; IAE: [PLATES]

1.3.2.1 NPL → NCO - PLZR; NOK; IAE: NCO + PLZR

1.3.2.1.1 NCO → [PLATE] @ BCL, OK

1.3.2.1.2 PLZR → ZERO, NOK; IAE: PLZR = +S, CLF

8.Ø [MY SISTER AND THEM] → [MY SISTERS]* [MY SISTER AND HER FRIENDS]

9.Ø TAG → [YES]

Glossary of keywords

ADV - adverb(ial), ANIM - animate, ASOC - associative

AUX - auxiliary, BCL - base form final cluster, C- clause

CLF - final cluster results from suffixation, CMNT - comment

CTN - completion, DET - determiner, DOBJ - direct object, GR - grammar

HDW - headword, INAN - inanimate, IND - indefinite, IOBJ - indirect object

LEX - lexical, MASC - masculine, MC - main clause, N - noun,

NCO - countable noun, NEQV - no equivalent, NEUT - netral, P - phrase

PADJ - possessive adjective, PATT - pattern, PL - plural, PLZR - pluralizer

PRED - predicator, PREP - preposition, PRMD - pre-head modifier,

PROG - progressive, RD - third person, SG - singular, SUBJ - subject,

TEMP - temporal, TM - time, TRAN - transitive, VBL - verbal,

VT - verb used transitively

* - alternative parse or meaning, @ - absence of..., [ ] enclose literals,

, - end of information set, , - minor separator.

Developing the Computer System

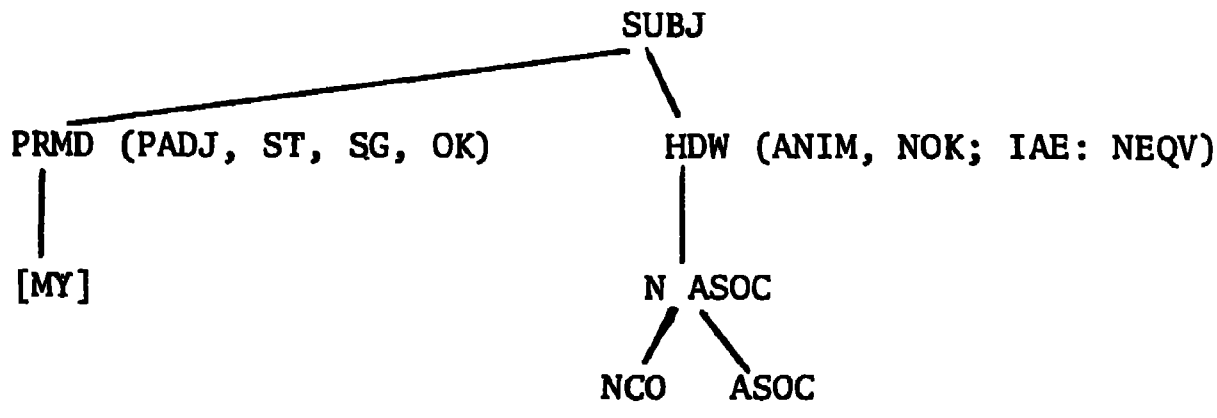
The structure of the parse tree is, in general, quite complex and a simple ad hoc approach to validity checking was quickly seen to be inadequate. As a result a formal description of the tree was developed and used to construct a (partially) syntax-driven validity checking routine. The output of this routine consists of a listing of the input, with error comments where necessary, together with the internal representation of the valid trees which is written onto a file - the parse-tree file - for the subsequent analyses.

Several other files are used in addition to the parse tree file. There is the informant file which contains profiles of the informants, (e.g. age, sex, linguistic background, etc), a set of form class files and a

set of classification files. The form class files are groupings of the various keywords which may occur in the data. Thus, for example, one form class file contains all keywords which may occur on the left-hand side of a rewrite. A classification file contains a group number for each informant; for example, one classification file contains 0 for each informant not aged 5 1 if the informant is aged 5 with a Hindi linguistic background and 2 otherwise. In any operation on the data the utterances of informants in group 0 of the relevant classification will be ignored.

Each node of a tree in the parse tree file consists of a name - in the case of a rewrite this is the left-hand side of the rewrite, otherwise it is the level number - and a set of descriptors, e.g. the grammar associated with the name. Thus, in the example of Figure 1, the lines 1.1, 1.1.1, 1.1.2, 1.1.2.1 become the sub-tree of Figure 2 where the descriptors are put in parentheses.

Figure 2



For any tree, each analysis starts at the root and many of the tasks to be described below may be regarded, in part, as a pattern matching exercise. The difficulties, and interest, arise because each node of the parse tree carries a substantial amount of information, and except for literals, only a partial matching of the nodes is usually required. In

addition, some tasks require the matching of disjoint sub-trees within a given parse tree, occasionally subject to side conditions which may involve nodes not lying on the paths between the root and any of the sub-trees of interest. Apart from the pattern matching,there is the problem of classification of the occurrences of the various patterns. This is a simple tabulation complicated, in some cases, by the fact that the total number of categories is unknown.

The basic task of the system may be cast in the form: count with respect to a given classification file, and subject to stated side conditions, the occurrences of a given pattern.

Since there are only 1,000 informants and they fall into a reasonably small number of classes it is economical to pre-classify on the basis of the informant profiles rather than build the classification process into the rest of the analysis. The system is instructed to produce a classification file by a statement of the form:

CLASS = ⟨ classification file name ⟩ , (⟨ expression list ⟩) where ⟨classification file name⟩ is the name by which the file will be known, and each expression in ⟨expression list ⟩ is a Boolean expression. For example:

CLASS = HIND1, (AGE = 5 ¢ LANG = HIND1, AGE = 5 ¢ LANG ≠ HIND1)

will produce the classification file given earlier as an example.

The side conditions refer to items in the parse trees which must occur if the tree is to be included in a given analysis. For example, if only affirmative active utterances are to be analysed the side condition Ø.7 AFM ACTV is used. The pattern to be used is stated in a manner similar to that used in specifying the input data. Thus, the pattern description

PRED → ... + AUX...; GR: @ CTN, NEUT TM, @ PROG, PATT

AUX → [DOES]

indicates that the sub-tree

PRED (GR: @ CTN, NEUT TM, @ PROG, PATT)
|
AUX
|
[DOES]

is of interest, subject to the convention that both the order of node descriptors (where given) and node descriptors not mentioned in the pattern are to be ignored. The occurrence of keyword FORM = < form class file name> indicates that the contents of the stated form class file are to form an additional dimension to the final tabulations. Thus the pattern

AUX → [?] FORM = OKFILE

where OKFILE contains the keywords OK and NOK and is an abbreviation for the pair of patterns.

AUX → [?] OK

AUX → [?] NOK

The symbol ? indicates that the items found there are also to add an additional dimension to the tabulations. The output of each tabulation may also be used to construct a classification file of the informants, to be used in further analyses.

CONCLUSION

In respect of performance of groups with different socio-linguistic descriptions, for purposes of this study, it is assumed that the frequency of occurrence of particular basic parse trees is a meaningful indicator of differences in speech patterns. A major difficulty is that no two trees in the study are identical but at the same time if we strip too much information

from each node there are too few trees to make an analysis worthwhile, and in part, the study aims at determining the degree to which stripping of information at interior nodes is necessary if the computer is to be a useful aid.

REFERENCES

Bickerton, D. 1973 — "The Nature of a Creole Continuum" Language 49 (3) p.640-669.

Carrington L. and Borely, C. 1969 — "An Investigation into English Language Learning and Teaching Problems in Trinidad and Tobago Progress Report". U.W.I. Institute of Education, St. Augustine (mimeo).

Carrington L., Borely, C. and Knight H. 1972 — Away Robin Run: A Critical description of the Teaching of Language Arts in the Primary Schools of Trinidad and Tobago. U.W.I. Institute of Education, St. Augustine. (mimeo).

Carrington L., Borely, C. and Knight H. 1974 — "Linguistic Exposure of Trinidadian Children" Caribbean Journal of Education No. 1, p.12-22.

De Camp, D. 1971 — "The study of pidgin and creole languages" in Hymes Pidginization of Creolization of Languages CUP. p.13-39.

Le Page, R.B. 1957 — "General outlines of creole English dialects in the British Caribbean". Orbes 6, p.373-391.

Knight, H., Carrington L. and Borely, C. 1974 — "Preliminary Comments on Language Arts Textbooks in use in the primary schools of Trinidad and Tobago". Caribbean Journal of Education No. 2 p.24-47.