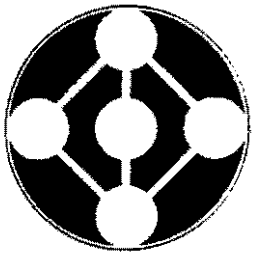

Alternative Technologies

# NATURAL LANGUAGE ALGORITHMS:

## A NEW SYSTEM OF GRAMMAR



## Introduction

Alternative Technologies and its associates have invested over twenty years of research into human communication. As a result, we have developed unique algorithms which describe the grammatical structures of natural languages. A portion of these algorthms have been programmed in such a way that they are portable. To date, the programs have been run on an IBM 370, a PDP-11, a Victor 9000, and an IBM-PC/XT.

We present here a brief description of the history and nature of these developments followed by a description of the work remaining to develop usable products.

## Background

The original purpose for the development of a new grammar was to develop a clear, logical description of the structural features of standard written English as an aid to the teaching of written composition. As time went on, it became increasingly evident that the system also was able to include structural variations characteristic of other levels of usage (including dialects) within one systematic description, rather than to dismiss such variations as examples of "non-standard" English.

Native speakers of English acquire their language resources in two categories: vocabulary and syntactic structures. Vocabulary acquisition comes first; words are learned - stored in the memory as "labels" for recognizable bits of experience. The earliest speech of the child is characterized by single words as isolated utterances. The acquisition of syntactic structures occurs next - at the point where two words are put together as an utterance in such a way that one of the words is used to say something about the other. Language acquisition then continues in both categories overlapping not only as simultaneous activities but also as mutually supportive: some words open new structural possibilities; some structures provide new vocabulary items.

Out of this stage of language learning there are four implications to underscore:

1. The basic language - vocabulary and structure - of the native speaker of English is acquired through experience.

2. The resources acquired at this time are limited primarily to the speech community of the child.

3. Structures acquired include the significant intonation patterns of the speech community.

4. Language resources acquired through this experience are regarded by the native speaker to be the norm; thus variations between his usage and "standard" English are considered to be the deviation of "standard" English from the norm.

It is entirely possible, then, that the lack of effectiveness in "teaching" grammar, particularly in developing ways of integrating the strudy of structure into writing improvement, is the result of grammars which are not only inaccurate but also incapable of describing variations in differing levels of usage, or in dialects, without denigrating the so-called sub-standard language practices or, at best, condescending to admit the propriety of such practices in culturally sub-standard situations.

The advantage of a grammar which identifies differences by encompassing those differences within one system is obvious. There are no value judgements stated, or even implied. The threat of loss of identity with his own community through change of language patterns can be largely removed from the student. All students can come to the study of structure in standard written English as a new "dialect" to be added to the language of their experience. Finally, such a grammar will be, by its very nature, extremely flexible and, therefore, powerful.

Such a grammar has, in fact been constructed. It is based upon three simple postulates which result in approximately one hundred (100) simple transition rules for the written English language. As lengthy, detailed, complicated, and perhaps tedious as an explanation of the grammar might be, it should be realized that the system is capable of generating (and analyzing) thousands of sentences from but one verb and three pronouns. Furthermore, the grammar is compatible with computer programming, since all

decision points are binary - involving yes/no choices - and obligatory word form instructions are weel-defined and incorporated into the structure. The cybernetic nature of the system has the additional advantage of offering integrated demonstrations of the process of language formation without requiring student or teacher to know the system of grammar. The system reinforces language rather than itself.

There are obvious similarities between this approach and that of generative/transformational theory. The differences, though less obvious, are significant:

1. The "tree-branch" diagram of the generative/ transformational theory suggests that language formation as a process moves from structure to meaning. Students who manipulate the formula for producing a kernel sentence are removed from language itself until the phrase structure rules are exhausted and lexical items begin to be inserted. The tansition rule approach, on the other hand, depends to a great extent on word choices for the gneration of phrase structure as well as morphology.

2. Transformational operations are awkward, theoretical, and, again, removed from the actual language process. They force student preoccupation with the system itself and preclude any reinforcement of the student's native intuition or feel for his own language. The student who works with the transitions rules, hovever, is encouraged to check the structures, as they form, with his own feeling about his language, to identify significant points where his own language usage differs from what is being revealed to him as he formulates his samples.

3. The transitions rules approach enables the student to "challenge the system". By creating different transition rules, by by-passing some obligatory choices, by substituting choices, he not only can see some of the non-English structures that are produced but may well discover the "folk analogy" bases for dialect or idiolect constructions.

4. The actual utilization of the transition rules to produce predications which are not included in the structural resources of the student is quite simple, differing from actual language formation only in the imposition of a formula before the actual first step, word choice.

In order for the system to work without exception, many of the precepts of traditional grammar, preserved in newer theories, have been discarded. First of all, all previous grammatical analysis has had to show "be" as an exception: no verb in English has more than five principle parts except "be", which has eight. It turns out that this can be resolved by somewhat altering the traditional concept of "agreement".

Similarly, the structural options for signaling tense do not necessarily conform to the concepts we hold of past, present, and future. It seems far better to suggest that, in any verb element, there is one alternate form which generally (but not always) indicates a change in time reference.

Finally, the many implications for comparative studies of dialect differences in Enlish cannot be covered here. However, as suggested above, the concept of dialects as "outside" the system is abandoned since structural distinctions are easily managed within the system.

Since the development of the system, numerous tests have been made regarding its correctness and its usefulness. As mentioned in the introduction, this system of grammar has resulted in new algorithms for parsing, correcting, and teaching the grammatical structures of natural languages. The first attempt to test these ideas resulted in the use of a toy train which students used to "build" sentences. As a teaching tool, the train was quite successful. A portion of the algorthms have been programmed in such a way that they are portable. To date, the programs have been run on an IBM 370, a PDP-11, a Victor 9000, and an IBM-PC/XT and can be used to generate hundreds of thousands of grammatically valid sentences from a small lexicon and an even smaller program.

Finally, the system has been used to teach a college-level course of English written composition to non-native speakers having English as their second language. Tradition might it likely that no more than two-thirds of the students would pass the standard composition examination: a full ninety-four percent passed the examination. The course received review comments from the students such as "I was able to feel comfortable with the rules for the first time".

## Proposed Project

The project involves a series of developments, each of which will result in a saleable software product. So far, the programs implement a portion of the algorithms (the portion for one of three primary types of sentence structure) which generates all the grammatically correct sentences possible from a selected group of words (up to a sentence of maximum length). These restrictions of sentence type, length, and lexicon choice are artificial, being useful for development and demonstration purposes. The program is small: three pages of source code in the C programming language.

The next phase (phase two) of our efforts is the alteration of the existing program from a generation function to a recognition or parsing function. We expect this phase to be completed by the end of August, 1885. The program will accept a sentence of the specified type, and flag grammatical errors.

Phase three of the project will address the type restriction, so that sentences of type two or three can be accepted as input. We expect this phase to be completed with six months of effort by two programmers and one linguist.

Phase four of the project will round out the package by removing the remaining restrictions on sentence length and lexicon choice. Also, during this time final product users documentation and marketing product descriptions will be produced. These efforts are expected to require two programmers and one linguist for two to four calendar months.

Each of these phases involves considerable documentation and testing effort. In addition to the personnel cited, a technical manager will supervise the project.

## Proposed Products

Several products are possible. A grammatical editor which searches for grammatical errors within computer text (for example, as an adjunct to a standard word processing system), would be most helpful in business and governmental applications, domestic and especially international. Ultimately, the editor might search the entire text in a batch mode, flagging all errors for correction. However, it is also just as likely that the final product could detect and partially correct errors as they occur. I say "partially correct" because a grammatical mistake frequently results from or at least indicates the possibility of ambiguous intent. For example "I might gone" could be corrected to say "I might have gone" or "I might go". The user must decide. The editor would also allow for "block edits" in which a tense or person was set or edited over a block of text of arbitrary length with a few keystrokes. We are of the opinion that it will take approximately a year of continuous effort to bring this particular product to a marketable form; namely, as a piece of software that can be integrated with an existing word processor and spelling corrector.

Another product that would be available much sooner would be an English grammar tutoring package. This piece of software would interactively allow the user to enter an arbitrary sentence, and would then request that the user alter the sentence in a specific way: for example, if "I am going" was entered, the program might prompt 'change to past tense, third person' or it might automatically change the 'I' to 'I have' ask the user to make a reasonable correction say by changing 'am' to 'to be'. The complexity of the drills can be increased arbitrarily. Such a package will be useful in schools teaching English communications, especially writing skills and English-as-a-second-language (ESL).

In later stages of research, we would like to duplicate the tools developed earlier for other languages. The algorithms are such that the choice of English as a natural language grammar to work on is a mere convenience for the linguists and programmers on the team at this time; nothing appears to restrict the algorithms applicability to other languages including German, French, and Japanese. It should also be noted that the algorithms can be made to handle dialects within a language. This would, of course, greatly enhance the markets for earlier products.

Finally, in the most advanced stages, we expect that the total group of algorithms will lead to better automatic translation tools. If the algorithms have been applied to the grammar of the source language, grammatical inconsistencies need not be improperly translated into the target language without correction. Furthermore, with automatic look-up restricted by the intented grammar of the source language as seen from the target language, many of the colloquialisms which are the bane of automatic translation attempts would be properly handled.

Sample Output

The attached page demonstrates the power of the algorithms. For the word choices shown, the sample program generates all possible grammatically correct English sentences, of which this page is a small portion. As pointed out above, the program is artificially limited by the programmer for reasons of practicality.

| | | | |
|---|---|---|---|
| I | may | go | |
| I | may | have to | go |
| I | may | start | going |
| I | may | have | gone |
| I | may | be | gone |
| I | may | be | going |
| I | do | go | |
| I | do | have to | go |
| I | do | start | going |
| I | go | | |
| I | have to | go | |
| I | start | going | |
| I | have | gone | |
| I | am | gone | |
| I | am | going | |
| I | am to | go | |
| I | might | go | |
| I | might | have to | go |
| I | might | start | going |
| I | might | have | gone |
| I | might | be | gone |
| I | might | be | going |
| I | did | go | |
| I | did | have to | go |
| I | did | start | going |
| I | went | | |
| I | had to | go | |
| I | started | going | |
| I | had | gone | |
| I | was | gone | |
| I | was | going | |
| I | was to | go | |
| You | might | go | |
| You | might | have to | go |
| You | might | start | going |
| You | might | have | gone |
| You | might | be | gone |
| You | might | be | going |
| You | did | go | |
| You | did | have to | go |
| You | did | start | going |
| You | went | | |
| You | had to | go | |
| You | started | going | |
| You | had | gone | |
| You | were | gone | |
| You | were | going | |
| You | were to | go | |
| You | may | go | |
| You | may | have to | go |
| You | may | start | going |
| You | may | have | gone |
| You | may | be | gone |
| You | may | be | going |

| | | | |
|---|---|---|---|
| You | do | go | |
| You | do | have to | go |
| You | do | start | going |
| You | go | | |
| You | have to | go | |
| You | start | going | |
| You | have | gone | |
| You | are | gone | |
| You | are | going | |
| You | are to | go | |
| We | do | go | |
| We | do | have to | go |
| We | do | start | going |
| We | go | | |
| We | have to | go | |
| We | start | going | |
| We | have | gone | |
| We | are | gone | |
| We | are | going | |
| We | are to | go | |
| We | may | go | |
| We | may | have to | go |
| We | may | start | going |
| We | may | have | gone |
| We | may | be | gone |
| We | may | be | going |
| We | might | go | |
| We | might | have to | go |
| We | might | start | going |
| We | might | have | gone |
| We | might | be | gone |
| We | might | be | going |
| We | did | go | |
| We | did | have to | go |
| We | did | start | going |
| We | went | | |
| We | had to | go | |
| We | started | going | |
| We | had | gone | |
| We | were | gone | |
| We | were | going | |
| We | were to | go | |