Sentence Structure Diagrams

Susumu Kuno

Computation Laboratory, Harvard University

A system for automatically producing a sentence structure diagram for each analysis of a given sentence has been added to the program of the multiple-path syntactic analyzer. A structure code, consisting of a series of structure symbols or phrase markers that identify the successive higher-order structures to which the word in question belongs, is assigned to each word of the sentence. The set of structure codes for the words of a given sentence is equivalent to an explicit tree diagram of the sentence structure, but more compact and easier to lay out on conventional printers.

The diagramming system makes some experimental assumptions about the dependencies of certain structures upon higher-level structures. All the major syntactic components of a sentence (i.e., subject, verb, object, complement, period, or question mark) are represented in the current system as occurring on the same level, all being dependent on the topmost level, "sentence". A floating structure such as a prepositional phrase or adverbial phrase or clause, whose dependency is not determined in the analyzer, is represented as depending upon the nearest preceding structure modifiable by such a floating structure. Different assumptions as to structural dependencies would yield different diagrams without requiring modification on the main flow of the diagramming program.

The diagrams thus obtained contribute greatly to the rapid and accurate evaluation of the analysis results, and they are also useful for obtaining basic syntactic patterns of analyzed structures, and for detecting the head of each identified structure.

Linguistic Structure and Machine Translation

Sydney M. Lamb

University of California, Berkeley

If one understands the nature of linguistic structure, one will know what design features an adequate machine translation system must have. To put it the other way around, it is futile to attempt the construction of a machine translation system without a knowledge of what the structure of language is like. This principle means that if someone wants to construct a machine translation system, the most important thing he must do is to understand the structure of language.

Any MT system, whether by conscious intention on the part of its creators or not, is based upon some view of the nature of linguistic structure. By making explicit the underlying theory for various MT systems which have been proposed we can determine whether or not they are adequate. Similarly, by observing linguistic phenomena we can determine what properties an adequate theory of language must have, and such determination will show what features an MT system must have in order to be adequate.

It can be shown that some of the approaches to MT now being pursued must necessarily fail because their underlying linguistic theories are inadequate to account for various well-known linguistic phenomena.

On Redundancy in Artificial Languages

W. P. Lehmann

Linguistics Research Center, The University of Texas

Artificial languages are one concern of work in computational linguistics, if only as a mnemonic device for interlinguas which will be developed. Even if it does not gain wider use, the structure of an artificial language is of general interest.

In contrast to the artificial languages which have been widely proposed, linguistic principles underlying a well-designed artificial language and its usefulness are well-established, particularly through Trubetzkoy's article, TCLP 8.5-21. which indicates phonological limitations for such a language. Since Trubetzkoy's specifications yield a total of approximately 11,000 morphemes, if an artificial language incorporated the degree of redundancy found in natural languages it would be severely handicapped by the size of its lexicon. The paper discusses the problem particularly with regard to suprasegmentals, which Trubetzkoy almost entirely ignored.

A Procedure for Automatic Sentence Structure Analysis

D. Lieberman

IBM Thomas }. Watson Research Center

The two main considerations in the design of this procedure were the economical recognition and representation of multiple readings of syntactically ambiguous sentences, and general applicability to "all" languages (English, Russian, Chinese). The following features will be discussed: types of structural descriptions, form of linguistic rules, use of linguistic heuristics to achieve economical multiple analyses, application to linguistic research and application to production MT systems. Also, the relation between this procedure and other existing sentence analysis procedures will be discussed.

An Algorithm for the Translation of Russian Inorganic-Chemistry Terms

L. R. Micklesen and P. H. Smith, Jr.

IBM Thomas J. Watson Research Center

An algorithm has been devised, and a computer program written, to translate certain recurring types of inorganic-chemistry terms from Russian to English. The terms arc all noun-phrases, and several different types of such phrases have been included in the program. Examples are:

AZOTNONATRIEVA4 SOL6	sodium nitrate
SOL6 ZAKISI/OKISI JELEZA	ferrous/ferric salt
ZAKISNA4 OKISNA4 SOL6 JELEZA	
GIDRAT ZAKISI/OKISI JELEZA	ferrous/ferric salt

etc., where the stems underlined may be replaced by any of a number of other stems (up to 65 in some positions) in the particular type.

Translation of each type encounters problems common to almost all the types: (1) The Russian noun is translated as an English adjective, while the noun of the resulting English phrase is found among the modifiers of the Russian noun. (2) The Russian noun (English adjective) may be a metal with more than one valence state, the state indicated (if at all) by the modifiers. (3) The number of the resulting English noun-phrase is determined by some member of the Russian phrase other than the noun. (4) The phrase elements may occur compounded in the chemical phrase but free in other contexts, and dictionary storage must provide for this. The program permits translation of conjoined phrase elements as well.

The paper also includes an investigation into the deeper grammatical implications of this type of chemical nomenclature, and some excursions into the semantic correlations involved.

The Application of Table Processing Concepts to the Sakai Translation Technique

A. Opler, R. Silverstone, Y. Saleh, M. Hildebran, and I. Slutzky

*Computer Usage Company**

In 1961, I. Sakai described a new technique for the mechanical translation of languages. The method utilizes large tables which contain the syntactic rules of the source and target languages.

As part of a study of the AN/GSQ-16 Lexical Processing Machine, a modification of the Sakai method was developed. Five of six planned table scanning phases were implemented and tested. Our translation system (1) converts input text to syntactic and semantic codes with a dictionary scan, (2) clears syntactic ambiguities where resolution by adjacent words is effective, (3) resolves residual syntactic ambiguities by determining the longest meaningful semantic unit, (4) reorders word sequence according to the rules of the target language and (5) produces the final target language translation.

French to English was the source-target pair selected for the study. An Input Dictionary of 3,000 French stems was prepared and 17,000 entries comprised the Input Product Table (allowable syntactic combinations).

Since Sakai was working with highly dissimilar languages, he found it necessary to use an intermediate language. Because of the structural similarity between French and English, we found an intermediate language was unnecessary.

The method proved straightforward to implement using the table lookup logic of the Lexical Processor. The translation was actually performed on an IBM 1401 which we programmed to simulate the concept of the AN/GSQ-16 Lexical Processor. In our implementation magnetic tapes replaced the photoscopic storage disk.

Slavic Languages—Comparative Morphosyntactic Research

Milos Pacak

Machine Translation Research Project, Georgetown University

An appropriate goal for present-day linguistics is the development of a general theory of relations between languages. One necessary requirement in the development of such a theory is the identification and classification of inflected forms in terms of their morphosyntactic properties in a set of presumably related languages.

According to Sapir, “all languages differ from one another, but certain ones differ far more than others”. As for the Slavic languages he might well have said that they are all alike, but some are more alike than others. The similarities stemming from their common origin and from subsequent parallel development enable us to group them into a number of more or less homogeneous types.

The experimental comparative research at The Georgetown University was focused on a group of four Slavic languages, namely, Russian, Czech, Polish and Serbocroatian.

The first step in the comparative procedure here described is the morphosyntactic analysis of each of the four languages individually. The analysis should be based on the complementary distribution of inflectional morphemes. The properties whose distribution must be determined are:

1) the graphemic shape of the inflectional morphemes,
2) the establishment of distributional classes and subclasses of stem morphemes and (on the basis of 1 and 2),
3) the morphosyntactic function of inflectional morphemes which is determined by the distributional subclass of the stem morpheme.

f(x,y)-l, where x is the distributional subclass of the stem morpheme (which is a constant) and y is the given inflectional morpheme (which is a free variable). On the basis of this preliminary analysis the patterns of absolute equivalence, partial equivalence, and absolute difference can be established for each class of inflected forms in each language under study.

Once this has been accomplished, the results can be used in order to determine the extent of distributional equivalences among the individual languages. The applicability of this procedure was tested on the class of adjectivals. Within the frame of adjectivals the follow-

* This work was performed while under contract to IBM Thomas J. Watson Research Center, Yorktown Heights, New York.