[From: *Studies in machine translation*, international workshop…Riyadh, March 1985]

Computer Aided Translation

State of the art in the USA and Canada

Paper prepared for the Saudi Arabian National Center for Science and Technology
International Workshop on CAT
Riyadh, March 16-19 1985

Winfred P. Lehmann, *The University of Texas*

Translation is a use of language. Like any such use, it may be performed intuitively or after specific training. Until our generation all translation had to be performed without the assistance of mechanical aids. Less than forty years ago several eminent scientists — non-linguists — proposed that the newly developed machine, labeled computer, might devote some of its capabilities to translation. That proposal by non-linguists may well have led to the notion that computerized translation is somehow more dependent on the understanding of computers and their workings than on the understanding of language. This notion must be totally rejected for an accurate view of computer aided translation (CAT), whether machine-aided or completely carried out by computer. Translation is today, as it always has been, a use of language. Success in that use depends on skills which in turn depend heavily on an accurate understanding of language.

A report on the state of the art must therefore focus on the current status of our understanding of language, more specifically, on an accurate understanding based on investigation of language in use, not on language of an 'ideal speaker-listener' in a speech community existing only as a figment of an academic mind. My report accordingly deals with the most advanced systems of machine translation in the Americas today, presented with perspective through reference to past approaches and to theoretical work carried out independently of machine.

Because of misunderstandings regarding computerized translation, a parallel from another activity may be useful to clarify the activity of translating. Such a parallel is transportation. Human beings have long possessed skills in transporting themselves from one place to another. Mechanical devices which have speeded up the process were developed somewhat before computers. Although these devices have brought advantages to the activity, the ends are still the same. Whether we transport ourselves from one spot to another on foot, by means of animals, or by mechanical means, through transportation we shift ourselves and our effects from one place to another, as human beings have from remote ages.

Clearly, any change in technology involves change in the procedures, in their mastery and in their control. When Akkadian linguists developed dictionaries five millennia ago, translators acquired capabilities which they previously lacked. But they also had to master the art of writing and the manipulation of clay tablets -- in modern terms, software and hardware. When human beings learned that sandals would permit less troublesome transportation, they had to master the art of preparing and fashioning leather. As new technology was introduced, transportation like translation acquired various advantages, but also the need to control the new technology. We must never forget however that the technology is only ancillary to the central procedure, whether transportation or translation. Computers used for translation have in no way changed the activity noted above, that is, one of the uses of language. They have simply changed the procedures involved, much as the dictionaries on clay tablets five millennia ago must have changed the procedures of the Akkadian translators.

This is not to deny the importance of the procedures. Actually, an examination of the computational procedures, the available hardware and software illuminates the history of translation involving computers and the attitudes regarding such translation.

An attempt to portray computation before the day of the transistor, let alone the microchip, and before the day of high-level programming languages like LISP encounters receptivity among today's young specialists comparable to portrayal of horse and buggy transportation. Yet those were the technological aids that early workers in CAT, like Edwin Reifler and Leon Dostert had to apply, not to mention workers in countries outside the scope of this review. It is small wonder that they introduced techniques which could be applied only to specific limits. What is remarkable is their success and their lasting contributions. Among the limiting techniques was concentration on lexical elements rather than syntactic patterns including lexical elements. Among lasting contributions is Reifler's insistence on sorting out for translation what we now call sub-languages rather than embracing all of language.

Concentration on lexical elements was not without success. This very week I had a conversation with one of the leading scientists at the Oak Ridge National Laboratories who still finds their version of the old Georgetown system useful. He however controls many languages, so that he recognizes when a given lexical item is reproduced with an equivalent that does not fit in the sub-language at hand. He also knows the syntactic structure of French, German and Russian, so that he can arrange the lexical elements in appropriate order, substitute syntactic devices that mark definiteness, such as articles in English, through these skills capturing a fair notion of the meaning and usefulness of the original. He also knows the limited purpose of lexical translation; it permits him to decide whe-

ther to request a more complete translation. For multilinguals, including professional translators, lexical translation is not without merit. For less skilled users, not to speak of the American tourists who bought pocket translators some years ago, lexical translation has few benefits. Further, in the current state of the art it has been superseded and is therefore of limited value.

However distasteful the reference, any review of the state of the art must mentioned the 1966 report of the American Academy of Sciences on machine translation, for which one can scarcely find a kinder adjective than unfortunate. Even though we would like to forget it — though not as vigorously as should the authors or the august Academy itself - we must recall that it was directed against a largely lexical approach with the primitive computers and software of the day, which the Pierce Committee declared to yield results out of keeping with expenditures — the total of which would not buy a single fighter plane. But for the Academy report, today's state of the art would be far advanced beyond the actual situation. Subsequent research was carried out with minimal support, yet also with some success.

The systems which have been developed in the Americas, and the extent of their development have been admirably described in a report by Dr. Jonathan Slocum (1984). Its ready availability obviates the need to present here an account of the work by the groups at Georgetown University, the University of Montreal and Brigham Young University. Similarly, the Slocum paper sketches the status of well-known systems like SYSTRAN, LOGOS, METEO, WEIDNER, SPANAM and ALPS, as well as systems outside the Americas. Work of other groups, such as those at IBM cannot be sketched; company policy is more concerned with secrecy than with general scientific advance. Besides reports like Dr. Slocum's, an account of the state of the art should be directed at the achieved level of understanding of human language applicable today in a CAT system.

For this purpose I outline the conception of language which human users command and which then is the goal in computational command of language, however partial at present.

Language, by a conception following Peirce, now more and more widely accepted, is a communication system with three relationships, in accordance with three items concerned in communication. These we may label the user, the sign, the universe, of which the sign is central.

Relationships between signs are studied in grammar, Peirce's syntactics.
Relationships between signs and the universe are studied in semantics.
Relationships between signs and users are studied in pragmatics.

While the topics involved in each of these areas have long been studied, formalized approaches were first applied in the late 19th century, by logicians. We leave open the question whether Panini's grammar of Sanskrit represents a formalization; acquaintance with it may at least have influenced linguists to formalize their approach. Formalization is essential for computational manipulation of language.

To understand the state of the art of CAT we must be aware of the current status of research in general linguistics on the one hand, in applied computational linguistics on the other. Since we are focussing here on applied computational linguistics, specifically CAT, we may first sketch its current status with reference to each of the three facets of language study.

Pragmatics, the least accounted for, is sidetracked in CAT by its restriction to scientific and technical language which requires depersonalized application by the user. In short, CAT has made an end-run around pragmatics by identifying sub-languages and focussing on them. I may mention that pragmatics is also the least understood area of language in current theoretical study.

Semantics, somewhat similarly, is largely evaded in MAT through pairing of roughly equivalent elements of language by human analysts. For example, an Arabic lexical element like kataba is paired with English write, German schreiben and so on rather than classified in an Arabic set, which set is then analyzed and described for semantic features to be represented formally in computational notation. Similarly, meaningful syntactic patterns like genitival phrases or clause patterns are essentially matched by language pairs rather than represented formally in some universal semantic framework. Computational attempts to develop such a framework are still in their infancy. Work in theoretical linguistics is similarly tentative.

Syntatics then is the only facet of language study which current computational procedures manage. Even here attempts at control are partial.

CAT evades the problem of controlling one segment of syntatics, phonology, by using written texts. I therefore pass over the large amount of work designed to achieve input and output of spoken language, even though this has some success to its credit.

Current CAT also fails to confront many morphological problems. It must control inflectional morphology, and in general does. Attempts to deal with derivational morphology, on the other hand, were pretty well dropped when large storage devices became available on computers. Thereupon it was more economical to deal with morphological items like compounds as units rather

than to analyze them for their constituents and then devise specific programs for their interpretation. Derivational morphology is also weakly controlled by general linguistics.

What is left is a stripped down form of language, consisting of lexical elements and their characteristic arrangements. Examples of such arrangements are clause patterns, like relative constructions with their markers, nominal modifying patterns, anaphoric devices. Restriction to lexical items and selected syntax may seem to involve a paltry fragment of language as a device for communication. But apart from its work on phonology, general theoretical linguistics has not achieved much more command of language, as we may now note.

When surveying pertinent work in general linguistics we may begin with Leonard Bloomfield's precepts for formalization, stated in his 'Set of Postulates for the Science of Language', LANGUAGE 2: 153-64 - 1926. The article restricts itself of precepts, as in declaring the 'postulation of zero elements ... necessary'. Most of Bloomfield's subsequent work with language presented his results discursively, not in formalized notation. Formalization of syntax in the Americas was introduced by Zellig S. Harris, in his paper 'From Morpheme to Utterance', LANGUAGE 22: 161-83 - 1946. His formalization is directly applicable in computational linguistics, with trivial modifications. As an example, BC = A is his equation 'to indicate substitutability', Rather than include further examples, I cite from his first footnote: 'In view of the fact that methods as mathematical as the one proposed here have not yet become accepted in linguistics, some apology is due for introducing this procedure. However, the advantage which may be gained in explicitness, and in comparability of morphologies, may offset the trouble of manipulating the symbols of this procedure. Furthermore, the proposed method does not involve new operations of analysis. It merely reduces to writing the techniques of substitution which every linguist uses as he works over his material. One works more efficiently when one thinks with pencil and paper'. Many today would substitute for the last four words: 'at a computer terminal'.

Two other steps towards formalization may be noted: that of Otto Jespersen in his ANALYTIC SYNTAX, Copenhagen: Munksgaard, 1937; that of Erwin Koschmieder, as in his BEITRAGE ZUR ALLGEMEINEN SYNTAX, Heidelberg: Winter, 1965, notably the essay of 1956: 'Die Mathematisierung der Sprachwisenschaft', pp. 124-39. The disruption of scholarship caused by WWII lessened their impact. Both, as well as references in Koschmieder, still merit study.

An early application of formalization in MAT in accordance with the Peircean conception of communication was directed by E.D. Pendergraft (MLTS 1961).

Following Carnap's terminology it distinguishes between formation [Formbestimmungen] and transformation rules [Umformungs-bestimmungen], which are rules of deduction (Carnap 1942:251). The transformation rules applied in the Linguistics Research Center system, comparable to those of Harris, are concerned with actual text rather than mental constructs. This is the only possible approach for successful computerization of language. While it contrasts sharply with Chomsky's concerns, it is increasingly observed in currently proposed linguistic theory as well as in computational linguistics.

Since Chomsky's attention to language dominated linguistic study during much of the last two decades, and still has many devotees, it requires note even though it has at best peripheral pertinence for computational linguistics. CAT can do little with grammars derived from the language of an 'ideal speaker-listener, in a completely homogeneous speech community, who knows its language perfectly' (Chomsky 1965:3). As is widely known, Chomsky's approach has been greatly modified, after transformations in the various versions of the Standard Theory were eliminated. Yet the aims remain unchanged. As Hans Bennis and Anneke Gross summarize it: 'The goal of linguistic theory is to provide an explanation for the language-facility — seen as a biological endowment — as reflected in the linguistic descriptions for any given language' (1980:8). And Chomsky lists three basic questions for his current approach (1984:11):

1. What constitutes knowledge language?
2. How is knowledge of language acquired?
3. How is knowledge of language put to use?

The problems have great fascination. But they have psychological rather than linguistic pertinence. For understanding language, whether for its own sake or for applications like those in computational linguistics it is difficult to evade Bloomfield's objection to attempts to explain one unknown in terms of another. Moreover, crucial examples of Chomsky's are hardly those to be dealt with in CAT, if in real life, such as the second below (1984:19):

John is too clever to expect us to catch Bill
John is too clever to expect us to catch

In spite of Chomsky's disclaimer, linguistics, and certainly computational linguistics, must deal with E-language (externalized language), which he characterized as language 'understood independently of the properties of the mind/ brain'. While it is true that a 'person's knowledge of language ... is not squarely addressed' by this approach, one cannot accept Chomsky's second objection, that it cannot account 'for the unbounded character of the E-language' (1984:29-30). Computational rules can indeed deal with this property of human language.

Happily, other current approaches attracting attention include tests of hypothetical constructs against facts of language, as indicated by Gazdar et al (GPSG) in the preface to their forthcoming book: 'This book contains a fairly complete exposition of a general theory of grammar.

Unlike much theoretical linguistics, it lays considerable stress on detailed specifications both of the theory and of the description of parts of English grammar that we use to illustrate the theory. ... One must set about some function [that assigns to each grammatical and meaningful sentence of English an appropriate structural interpretation], or one is not in the business of theoretical linguistics.' The sentiment is welcome, both with reference to general linguistics and to computational linguistics. Unfortunately the parochiality of generative grammar has not been overcome in restricting the scrutiny of language to English; apart from the distasteful ethnocentrism, such restriction runs great danger in leading to conclusions based on 'description of ... grammar' of only one language.

Moreover, like other works in generative linguistics GPSG deals with only 'parts of... grammar'. A CAT system cannot confine itself to 'parts of grammar', but must include rules for all grammatical patterns in use.

Another widely examined approach is lexical-functional grammar (LEG), developed by Kaplan and Bresnan. In the Chomskyan tradition it is based on the assumption 'that an explanatory model of human language performance will incorporate a theoretically justified representation of the native speaker's linguistic knowledge (a *grammar)'*. The chapter including the LEG 'formal system for grammatical representation ... presents a formalism for representing the native speaker's syntactic knowledge (1984:173). The formalism is said to 'have been designed to serve as a medium for expressing and explaining generalizations about the syntax of human language' (208). Moreover, it is asserted to be 'a restricted, mathematically translatable notation for which simple, psychologically plausible processing mechanisms can be defined'. (173-74). LEG may be useful for CAT, as in setting up a lexical component in addition to the 'syntactic, semantic, and phonological components of a grammar' (175). Similarly important will be grammars of languages differing from the SVO structure of English, such as VSO Arabic and OV Japanese.

Reference to the current work of Chomsky, Gazard, Kaplan/Bresnan and their associates does not exhaust the activities of general linguistic study in the Americas. Postal and colleagues have their own approach, arc-pair grammar. Sydney Lamb, drawing on the earlier theory of Louis Hjelmslev, is continuing his attention to stratificational grammar. The tagmemic theory of Kenneth Pike is still in use. Journals, such as the forthcoming issue of COMPUTERS AND

THE HUMANITIES, and conferences on computational linguistics provide amplification of these and even other approaches, as their publications and the copious bibliographies in the works cited here indicate.

Moving from explorations towards developing further approaches, we welcome the increased attention to discourse or text, an approach advancing beyond grammars confined to the treatment of sentences. Theoretical study has scarcely advanced the identification of characteristic features, which Beaugrande labels standards. Among these are *cohesion,* referring to syntactic devices, *coherence,* referring to semantic entities, *situationality,* referring to features fitting appropriately the topic concerned; further standards of *intentionality availability, informativity, intertextuality* involve speaker and hearer, requiring attention to pragmatic problems sidetracked by concentration on technical and scientific language in CAT. It should not be surprising to learn that text linguistic study is in its infancy, in spite of enormous publication involving specialists in sociology, anthropology, psychology, philosophy, communication as well as in linguistics. Treatment of texts is also central in artificial intelligence (AI).

Yet some scholars have outlined models dealing with texts as macro-structures. Beaugrande/Dressler propose the Augmented Transitional Network (ATN) for texts as well as for use in sentences, as noted by Slocum (1984:9); processing of sentences and texts might then be similar. Yet treatment of texts rather than sentences has led some scholars to focussing on meaning rather than on form, as Mel'cuk's text linguistic model may illustrate. By it a text is held to maintain connectivity through units of meaning, not in the first instance through syntactic units. The units of meaning are located in a deep lexicon, elements of which in the production of a text are put together by means of a deep syntax. As the characterization 'deep' suggests, both lexicon and syntax are more abstract than the lexicon and syntax of a grammatical approach, including that of computational linguistics (Beaugrande/Dressler 1981:27-29).

In view of the highly theoretical units and relationships proposed in such text linguistic approaches, it is hardly surprising that they provide no examples of complete descriptions of a text which might then be treated computationally.

Yet a similar approach has attracted considerable attention in the activities of Roger Schank and followers. The sphere of these activities is not the sentence, nor the text, but knowledge. Assuming for the present the common sense view of knowledge, we may only note that Schank and his colleagues do not equate it with 'language expressions that represent or convey it' (Beaugrande/Dressler 1981:85). Instead, they deal with situations like a birthday party, or a restaurant scene, about which speakers (presumably of a specific culture) are assumed to

In view of this activity it may be useful to cite constructs which are supposed to grasp knowledge, that is, knowledge of given situations. These constructs may be stored in computers. *Frames* are the patterns of knowledge concerning a given situation, the items that are involved in it. *Schemes* present sequences ordered in time and by the causes for given situations. *Scripts* are standard plans assumed of the activities of participants in a given situation. When one contemplates virtually any text, even one as simple as a folktale or an account of a birthday party, the difficulty of using this approach scarcely needs comment.

Yet the approach has been proposed for CAT. (I omit here uses which may be carried out in expert systems as well as other routinized analyses of restricted situations which do not involve translation.) Arguments in favor of a knowledge-based approach were drawn from the conclusion that a computer equipped with a grammar and a lexicon encountered problems because it could not evaluate context (Beaugrande/Dressler 1981:216, with reference to Wilks and Schank). Moreover, improved translation on the basis of programming 'knowledge of the world ... [and of] *all* language operations' was assumed to be 'worth the high costs.' The visionary hopes may be realized in some future century. Today even proponents of the approach display doubts, as in the concluding session of the AAAI this last summer, which discussed the possibility of a negative reaction to all work in AI comparable to the ALPAC report of 1966. For the present it is comforting to recall that restriction to specific sublanguages yields highly valued translation. Moreover, such restriction even evades the problem of semantic analysis, relying almost entirely on a syntactic and lexical approach.

In view of that success I will not review the work carried out in formal semantics. Like work in formal syntax, this owes much to Rudolf Carnap (1942) as well as to Alfred Tarski. Subsequently associated with the name of Richard Montague, it is now widely pursued though not applied in CAT, and accordingly passed over here.

Work in pragmatics is even less developed. A general introduction by Levison is useful in guiding scholarship. A promising treatment by Tamly Givon is also under way. Control of pragmatics will be essential if one hopes to present texts to computers in such a way that they must determine the context of those texts. Even unlimited funding would make such a prospect dubious.

The shortcomings of much research in general linguistics for devising a CAT system in keeping with their approach do not negate gains which computational linguists may take from them, nor insights into language. Especially as we move beyond the closely related languages of Europe - equated by Whorf as one language labeled Standard Average European (SAE) - we profit greatly from work in universal grammar, such as that resulting from the Stanford project (Greenberg 1978) or articles like that of Hooper and Thompson on 'The discourse basis for lexical categories in universal grammar' (1984). Any CAT project must be thoroughly informed of current work in general linguistics.

Further, my insistence that understanding of language is the primary requisite for capable computerized control of language is not meant to deny the importance of skillful use of software and hardware, and awareness of advances in these areas. The 1984 paper by Slocum and colleagues provides adequate evidence of essential role of those tools, especially in achieving economical results.

In conclusion, we may compare the state of the art in CAT with that in another field, theoretical chemistry; in chemistry, unlike linguistics, theoretical is equated with computational. An article in the 22 February 1985 issue of SCIENCE: Theoretical Chemistry Comes Alive: Full Partners with Experiment' by W. A. Goddard III contains examples of computational study which provided insights beyond those achieved in the time-honored approaches in chemical research. In commenting on these insights Goddard states: 'One indication of the present state of modern theory is that, when faced with disagreement between theory and experiment, the theorists were sufficiently confident of their results that they continued to examine possible reinterpretation of the experiments until they stumbled onto the key idea' (921). Linguistics is far less developed than is chemistry; we cannot claim a 'first-principles explanation' for all of our field such as that chemists find in quantum mechanics from the 1920s. Yet it is useful to point out the success of 'computational' chemistry in a workshop dealing with computational linguistics of today.

Moreover, Goddard credits the gains of theoretical chemistry to advances in hardware as well as to advances in theoretical methods (921). We are aware of the close relationship between advances in CAT and in computers appropriate for dealing with language as opposed to number systems. Among the needs of CAT is large storage space and efficient retrieval from those stores. This same issue of SCIENCE includes a report by Gina Kolata: 'Changing Bits to Magnetic Blips' (932-33). The report concerns efficient storing of data on computer disks based on 'a highly theoretical result for the field of dynamical systems', a very abstract field of mathematics. This advance in hardware, like others, cannot fail to be important for CAT.

The sponsors of the Workshop should not then confine their attention to the current state of the art of CAT. Although MT was suggested four decades ago, only the software and hardware developments of the past five years have made its goals realistic. Those goals are among the simplest of those attainable in computerized control of language. In presenting the current state of the art of CAT, the Workshop with its sponsors should not close their eyes to the promises of the future.

Note: In keeping with the title of the Workshop I use CAT as a cover term for all translation involving computers, whether Machine Aided Translation (MAT) or Machine Translation (MT).

I express my gratitude to Carl Weir of the Linguistics Research Center for making available to me recent writings in general linguistics.

References

Beaugrande, Robert de and **Wolfgang Dressler,** 1981. Introduction to Text Linguistics, London: Longman.

Bennis, Hans and **Anneke Gross.** 1980. The Government-Binding Theory: an Overview. GLOW Newsletter. Subsequently published in Lingua e Stile XV, 1981.

Bloomfield, Leonard. 1926. A Set of Postulates for the Science of Language. Language 2: 153-64.

Carnap, Rudolf. 1942. Introduction of Semantics. Cambridge: Cambridge University Press.

Chomsky, Noam. 1965. Aspects of the Theory of Syntax. Cambridge: MIT Press.

- Oct, 1984 (ms). Knowledge of Language. Its Nature, Origin and Use. Cambridge.

Gazdar, Gerald, Ewan Klein, Geoffrey K. Pullum, Ivan A. Sag. 1984 (ms.) Generalized Phrase Structure Grammar.

Givon, **Talmy**. 1984 (ms.) Mind, Code and Context: Essays in Pragmatics.

Goddard, William A. III. 1985. Theoretical Chemistry Comes Alive. Full Partner with Experiment. Science 227: 917-23.

Greenberg, Joseph K. ed. 1978. Universals of Human Language. 4 vols. Stanford: Stanford University Press.

Harris, Zellig S. 1946. From Morpheme to Utterance. Language. 22:161-83.

Hopper, Paul J. and S.A. Thompson. 1984. The discourse basis for lexical categories. Language 60: 703-52.

Jespersen, Otto. 1973. Analytic Syntax. Copenhagen: Munksgaard.

Kolata, Gina, 1985. Changing Bits to Magnetic Blips. Science 227: 932-33.

Koschmieder, Erwin. 1961. Beiträge zur allgemeinen Syntax. Heidelberg: Winter.

Languages and **Machines:** Computers in Translation and Linguistics. 1966. A report by the Automatic language Processing Advisory Committee (ALPAC). National Academy of Sciences. Washington, D.C.

Lehmann, Winfred P. and E.D. Pendergraft. 1961. Machine Language Translation Study No. 9 Austin: The University of Texas.

- and **Rolf Stachowitz.** 1971. Feasibility Study of Fully Automatic and Quality Translation. Austin: The University of Texas. Also: Griffis AF Base: RADC TR-71-295.

Levinson, Stephen C. 1983. Pragmatics. Cambridge: Cambridge University Press.

Slocum, Jonathan. 1984. Machine Translation: its History, Current Status, and Future Prospects. Austin: LRC Working Paper. LRC-84-3.
- and colleagues. 1984. METAL: The LRC Machine Translation System. Austin: LRC Working Paper. LRC-84.2.

Winfred P. Lehmann, Director
Linguistics Research Center
The University of Texas
Austin, TX 78713-7247, USA.