MT LINGUISTICS AND MT LEXICOGRAPHY

AT THE UNIVERSITY OF WASHINGTON*

BY

ERWIN REIFLER

I. Introduction

MT research at the University of Washington was, as elsewhere in this country, sparked by Dr. Warren Weaver's Memorandum of July 15, 1949, and actually dates from November of that year. Very early in this research it became apparent that the most advanced methods and insights of modern linguistics are of great importance, but by no means sufficient to solve the linguistic problems with which we are confronted in MT. The reasons responsible for this insufficiency are the following:

(1) Modern linguists either did not feel the need for, never intended or did not have the time to search for and supply that amount of analytic and descriptive detail which, while adding nothing to the capabilities of qualified human operators of language, is of paramount importance for the mechanization of the translation process.

(2) Modern synchronic linguistics--and it is synchronic linguistics with which we are concerned here--largely limits itself to a study of languages in isolation, avoiding rather than encouraging a comparison between languages or a consideration of the phenomena of one language in the light of the phenomena of another.

(3) An important and influential section of modern structural linguists would, as far as this is possible, rather exclude considerations of meaning in their linguistic work.

(4) Modern synchronic linguists have, apart from their purely academic interests, the practical aim of improving the teaching of foreign languages rather than of the native language. Their pupils are not only animate, but also human operators, and before these begin their study, they know already at least one language, namely their native language. A knowledge of mathematics is for them not absolutely necessary although it has recently become an asset in linguistic studies.

MT linguistics, on the other hand, is faced with the problem of teaching and improving the teaching of languages to inanimate operators, machines. These pupils are, at present at least, highly accomplished mathematicians, better mathematicians, in fact, than any human can ever hope to be. But before they become our pupils, they know no language at all--that is, if we ignore the language of mathematics in this context. Consequently, we are in MT concerned not only with the teaching of at least one foreign language, but with the teaching of the native language as well. This special situation requires the following modifications in our linguistic studies:

Before a human being learns enough of two languages so that he is able to translate correctly from one into the other, he possesses already other knowledge and abilities without which such a translation would be impossible. It is still an open question whether we shall ever be able to embody that other knowledge and those other abilities in a machine. If we were able to do so today, our linguistic task in MT would be relatively simple. We could in our MT research limit ourselves to applying the methods devised by modern structural linguists for the teachers and learners of language.

It is, however, a fact that we are not yet able today to impart such additional knowledge and abilities to machines. Consequently we are justified in looking for additional ways and means, even unorthodox ones, if these can help our task. We have to extend the purview of our linguistic interests, limiting ourselves neither to the exclusive consideration of each language in isolation nor to the speech noises and their recorded form. Our primary aim is not merely the collection and description of linguistic facts of isolated languages, but the best possible representation in one or more languages of meaning expressed in another language. Linguistic phenomena of one language have to be considered in the light of the linguistic phenomena of another language, and, most important, the determination (pinpointing) of meaning has to be the paramount aim to which everything else is to be subordinated. With this aim before us, we have to intensify our research in order to obtain greater detail in morphology and syntax and in their capabilities of determining the grammatical and nongrammatical meaning of contextual units. We shall sometimes even have to sacrifice academic truth in favor of practicality.[1]

---

*Presented in a paper read before the International Conference for Standards on a Common Language for Machine Searching and Translation in Cleveland, Ohio, September 6-12, 1959. The paper will be published in the Proceedings of the Conference by Interscience Publishers, New York.

[1]Cf. "MT Linguistics," pp. 137-140, in my paper, "The Mechanical Determination of Meaning," Chapter 9 of Machine Translation of Languages, The Technology Press of the Massachusetts Institute of Technology and John Wiley & Sons, Inc., New York, 1955.

II. Lexicography

The University of Washington Machine Translation Project in June, 1959 completed the lexicographical phase of the research and prepared a Russian-English MT-operational lexicon of 170,563 entries on 556,141 IBM punch cards.[2] In this lexicon not only individual free and bound forms are treated as lexical units, but also a number of uninterrupted idiomatic free-form sequences. For the purposes of MT I had to make here a clear distinction between two kinds of idioms:

(1) Those that are idiomatic not only in terms of the source language, but in terms of the target language as well and therefore do not permit a word-for-word translation. An example is English the man is an ass. If Chinese is the target language, a word-for-word translation would result in an unintended and impossible meaning because the Chinese do not use the word for ass to denote a stupid person.

(2) Those that are idiomatic in terms of the source language, but, nevertheless, permit a word-for-word rendering in a target language because such a procedure results in an accurate idiomatic translation. In such cases the target language happens to share the idiom. This is, for example, the case with English the man is an ass if German is the target language. Der Mann ist ein Esel is the word-for-word translation as well as the correct idiomatic translation.

It is clear that such shared idioms do not present any problem in MT since they permit a word-for-word translation. On the other hand, whenever a source language idiom is not shared by the target language, we have to treat it as a single lexical unit and code it as such in the MT memory if we want the translation system to supply an idiomatic translation. Such a treatment of unshared idioms is at present only possible in the case of those which are not interrupted by contextual constituents not forming an integral part of the idiom.

Our distinction between idioms which do and those which do not permit a word-for-word translation is quite legitimate. Another distinction, however, which I had to make because it served a good practical purpose in our MT research at the University of Washington can hardly lay claim to academic legitimacy. This is the following distinction:

(1) Genuine Idioms. These are semantic units consisting of more than one free form which can not be translated word-for-word because such a translation is either completely unintelligible, or intelligible only in a sense not intended by the original author. The man is an ass in the case of the Chinese language as the target language is a good example for the first. An example for the second is German aufsitzen lassen which may mean to let mount, but is often used in the idiomatic sense of to leave in the lurch, to fool.

(2) Pseudo-Idioms. These are semantic units consisting of uninterrupted free-form sequences of relative great frequency which, strictly speaking, are not idiomatic in terms of the source language and, if translated word-for-word, are, in their translated context, often quite intelligible in the sense intended by the original author. And yet, many advantages accrue to MT if such non-idiomatic uninterrupted semantic units are treated as if they were genuine idioms. If they are namely not treated as if they were genuine idioms, if they are translated word-for-word, then their constituent forms would often have to be represented by more than one target equivalent. An example is English League of Nations whose word-for-word translation into German would be something like Liga/Bund/Bündnis von/an/aus/über/während/etc., etc. Nationen/Völkern. This would already be correctly intelligible, but is certainly a far cry from a translation into conventional German. If, on the other hand, we treat the English expression as if it were an idiom and code its three constituent free forms as a single lexical unit into the memory device of the translation system, the latter is able to supply the best possible translation, namely the appropriate idiomatic German translation Völkerbund, thus reducing drastically the number of target alternatives from something like ten to two.

This treatment of non-idiomatic sequences as if they were idioms is, of course, something contrary to the principles of academic linguistics. But it is justified in MT because it results in a better translation output.

Such pseudo-idioms are in all modern languages of civilized peoples extremely numerous because they denote important and frequently discussed things and ideas the overwhelming majority of which is in most languages denoted by expressions consisting of more than one free form. A large number of these pseudo-idioms is made up of technical terms which are usually accessible in specialized dictionaries. But very many are not technical terms and they are not found in any dictionary because everybody who knows the grammar of the language concerned can understand them if he knows, or looks up in a dictionary, the meaning or meanings of the constituent free forms of the semantic unit. These semantic units would have to be collected in a laborious search through large quantities of textual material. But the collection even of those which are important for the automatic translation of scientific publications could be costly in terms of time and money.

There is, however, a relatively quick, easy and cheap solution to this problem, but it is again one which may cause some academic shudders. This easy solution is based on a consideration of the content rather than the form of the semantic units concerned.

It is well known that the German language has an extraordinary tendency and capability of forming compound words. These composites fall into two groups, namely those already well established in the language and those created every day for the requirements of the moment. It is the first group in which we are interested here because they are already recorded in dictionaries and are thus readily accessible. Now all these well

[2] Cf. Linguistic and Engineering Studies in Automatic Language Translation of Scientific Russian into English, University of Washington, 1958.

established single-free-form compounds denote important and frequently discussed things and ideas which in other languages are mostly denoted by expressions consisting of more than one free form. If we make use of this conceptual experience stored in the German language, and collect all non-German equivalents of the non-technical high-frequency concepts expressed by the German single-free-form substantive compounds, we shall obtain a large and, very likely, very important number of uninterrupted semantic units which we can, in our MT lexicography, treat as if they were idioms, and we shall obtain them with comparatively little effort--that is, without the necessity of a money and time consuming search through large quantities of publications.

An example for such a German substantive compound is the previously quoted Völkerbund, the equivalent of English League of Nations. Other examples for such high-frequency concepts belonging to the general rather than the technical language are those denoted by:[3]

TABLE A

1.

German: Denkarbeit;

English: effort of thinking/work of the mind;

French: travail de tête;

Russian: a) работа ума b) умственное занятие c) усилие мысли;

Chinese: 心思之活動／腦力;

2.

German: Denkart;

English: manner of thinking/disposition of mind;

French: manière de penser;

Russian a) образ мыслей b) образ мышления c) способ мышления;

Chinese: 想法;

3.

German: Denkblatt;

English: memorial leaf/lines in remembrance;

French: feuille commémorative/lignes commémoratives;

Russian: a) посвятительная страница b) мемориальная страница

Chinese: 記念頁;

4.

German: Denkfähigkeit;

English: faculty of thinking/power of thinking;

French: faculté de penser;

Russian: способность мышления;

Chinese: 思考能力;

5.

German: Denkfaulheit;

English: slowness of thought/mental inertness;

[3] The order is the alphabetic sequence of the German equivalents as they occur in the large edition of Muret-Sanders Enzyklopaedisches Englisch-Deutsches Woerterbuch.

French: paresse d' esprit;

Russian: a)медлительность мысли b)леность мышления c)инертность мышления;

Chinese: 思想之遲鈍;

6.

German: Denkform;

English: mode of thinking;

French: façon de penser;

Russian: a)образ мысле b)форма мышления

Chinese: 思考方式;

7.

German: Denkfreiheit;

English: freedom of thought/liberty of opinion;

French: liberté de penser;

Russian: a)свобода мышления b)свобода мнения;

Chinese: 思想之自由;

8.

German: Denkgesetze;

English: laws of thought/laws of the mind;

French: lois de la pensée;

Russian: законы мышления;

Chinese: 論理學之法則/思考原理;

9.

German: Denkmünze;

English: commemorative medal;

French: médaille commémorative;

Russian: a)монета в память.... b)медаль в память... c)медаль;

Chinese: 記念章;

10.

German: Denkübung;

English: intellectual exercise/mental exercise;

French: exercice intellectuel;

Russian: a)интеллектуальное упражнение b)умственное упражнение

Chinese: 思考之訓練/思考之練習;

The German equivalents of the ten concepts exemplified in this table are all single-free-form compounds. The English, French, Russian (with the single exception of медаль in No. 9) and Chinese equivalents, on

the other hand, consist each of more than one free graphic[4] form. If in our MT lexicography we only treat their constituent free forms as lexical units--that is, if we do not use the conceptual approach--then most of these constituent free forms will in the automatic output be represented by more than one target equivalent. The following table, elaborated in consideration of the operational information stored in the Russian-English MT lexicon prepared by the University of Washington Machine Translation Project, will demonstrate this with the Russian examples in the preceding table:

TABLE B

Source Language	Target Language
1. a) работа ума	work(of)mind/wit
b) умственное занятие	mental occupation/study
c) усилие мысли	stress/effort (of)(to/for)thought(s)
2. a) образ мыслей	form/way/image (of)thoughts
b) образ мышления	form/way/image (of)thinking/thought(s)
c) способ мышления	method (of)thinking thought(s)
3. a) посвятительная страница	dedicatory page
b) мемориальная страница	memorial page
4. a) способность мышления	ability (of)thinking/thought(s)
5. a) медлительность мысли	sluggishness (of)(to/for)thought(s)
b) леность мышления	laziness (of)thinking/thought(s)
c) инертность мышления	inertness (of)thinking/thought(s)
6. a) образ мыслей	form/way/image (of)(to/for)thought(s)
b) форма мышления	(uni)form (of)thinking/thought(s)
7. a) свобода мышления	freedom (of)thinking/thought(s)
b) свобода мнения	freedom (of)opinion(s)
8. a) законы мышления	laws (of)thinking/thought(s)
9. a) монета в память	coin in/to/at/on/of/like memory
b) медаль в память	medal in/to/at/on/of/like memory
10. a) интеллектуальное упражнение	intellectual exercise
b) умственное упражнение	mental exercise

If, on the other hand, we use the conceptual approach, that is, collect all Russian, English, French, Chinese, etc., equivalents of the concepts expressed by the German single-free-form substantive compounds and treat them in our MT lexicography as if they were idioms, the automatic system will supply idiomatic translations for them. The following table exemplifies these idiomatic translations, using again the Russian expressions of the preceding two tables:

TABLE C

Source Language	Target Language
1. a) работа ума	work *of* the mind

[4] In consideration of the divergent problems of the Chinese language we have to speak here of "free graphic form" rather than merely of "free form."

	b)	умственное занятие	mental work
	c)	усилие мысли	effort of thinking
2.	a)	образ мыслей	manner of thinking
	b)	образ мышления	manner of thinking
	c)	способ мышления	manner of thinking
3.	a)	посвятительная страница	dedicatory page
	b)	мемориальная страница	memorial page
4.	a)	способность мышления	faculty of thinking/power of thinking
5.	a)	медлительность мысли	slowness of thought
	b)	леность мышления	mental laziness
	c)	инертность мышления	mental inertness
6.	a)	образ мыслей	mode of thinking
	b)	форма мышления	form of thinking
7.	a)	свобода мышления	freedom of thought
	b)	свбода мнения	liberty of opinion
8.	a)	законы мышления	laws of thought/laws of the mind
9.	a)	монета в память...	coin commemorating...
	b)	медаль в память...	medal commemorating...
10.	a)	интеллектуальное упражнение	intellectual exercise
	b)	умственное упражнение	mental exercise

It is easy to multiply such examples. If we treat every uninterrupted non-German source language phrase whose concept is in German expressed by a single free form as a single lexical unit, we can make sure of an idiomatic machine translation of the phrase and thus avoid a large amount of the superfluous clutter which characterizes word-for-word translations.

In our MT lexicography we treat as lexical units not only bound forms, single free forms which can not be inflected, and the headwords (nominative singular, present infinitive) of those which can be inflected, but also all paradigmatic forms of the latter. I extended, moreover, the distinction between non-paradigmatic and paradigmatic semantic units to include the idioms which I divide into non-paradigmatic and paradigmatic idiomatic sequences. I have to stress here that in MT we are not interested in monolingual idioms, but in bilingual ones, i.e., free-form sequences of the source language which are idiomatic in terms of the target language and therefore should not be translated word-for-word. An example for a non-paradigmatic idiom is English first of all (it can neither be declined nor conjugated). An example for a paradigmatic idiom is English to bark up the wrong tree (I, you, he, she, it, we, you, they, am, are, were, will be barking, barked, have barked, shall bark, the wrong trees, etc., etc.). In the case of such paradigmatic idioms it is necessary, of course, to include all their possible variations in the MT-operational lexicon in order to make sure of an idiomatic translation of the idiom under all circumstances.

Another interesting problem MT has to face is that of semantic units of the source language not included in the MT-operational lexicon. I distinguish here two kinds:

(1) Semantic units of the present source language either inadvertently or purposely omitted from the MT-memory. To deal with this problem effectively, the automatic translation system can be so designed that all source forms not identified in the machine memory are automatically printed out in red print in the target text in the sequence of the input text, and, where necessary, in an English alphabetization. This is actually the procedure envisaged for the University of Washington Machine Translation Project.

(2)Semantic units of the future source language whose constituents are already known and are, in fact, lexical units included in the MT-operational memory. These are extemporized compounds--that is, compound forms created for the requirements of the moment and therefore not included in any dictionary and unpredictable. In order to enable an automatic translation system to identify such compounds and to translate them, the compounding principles of the source language and the types of possible compounds have to be studied and procedures

developed for the automatic determination of the inner boundaries of the constituents of these compounds. These problems I have previously discussed in detail elsewhere.[5]

III. Morphology and Syntax

The problem of extemporized compounds belongs, of course, to the field of morphology. My research revealed that in any language which forms substantive compounds only thirty types are theoretically possible. Of these, however, only ten types make sense, and these are therefore the only linguistically possible types. I showed that these ten types present only four possible matching situations with which the design engineer has to deal. In nine out of these ten types the translation mechanism will be able to make a unique decision, whereas only in the case of one type of substantive compounds will it have to supply a double answer.

It is evident that the paradigmatic idioms and the paradigmatic form classes of individual free forms present problems of both morphology and syntax. Many of the latter are in a number of languages non-distinctive and therefore in isolation grammatically ambiguous. Very representative examples are German der, die, das, and den. But if we consider their environment and their syntactic relation to this environment then they are in most cases not ambiguous at all.[6] Research aiming at an automatic solution of this problem has been carried on at the University of Washington for some time.[7]

These problems concern the syntax of the source language alone. Another problem of great importance in MT is that of disagreements in the word order of the two languages concerned in the translation process. Also here we have to consider environmental factors in both languages if we want to elaborate the linguistic prerequisites for an automatic reshuffling of the word order of the source language text into that required by the conventions of the target language.

[5](a) "Mechanical Determination of the Constituents of German Substantive Compounds," Mechanical Translation, Vol. II, No. 1, M. I. T., July, 1955, pp. 3-14.
(b) "The Mechanical Determination of Meaning." (See footnote 1), pp. 144-48.

[6]Cf. my "The Mechanical Determination of Meaning" (see footnote 1), pp. 148-154.

[7]See Robert E. Wall, Jr. and Udo K. Niehaus, Russian to English Machine Translation with Simple Logical Processing, AIEE, Paper No. 57-1062, August, 1957.