[*Proceedings of the National Symposium on Machine Translation*, UCLA February 1960]

Session 7: THE DICTIONARY

THE SOLUTION OF MT LINGUISTIC PROBLEMS THROUGH LEXICOGRAPHY

Erwin Reifler
University of Washington

I believe it is appropriate to quote here a genuine statement of Confucius that is very applicable to machine translation. Confucius once said:

> "The gentleman seeks perfection in himself, the small man seeks it in others. "

Mutatis mutandis this means:

> "The gentleman seeks perfection in a linguist, the small man seeks perfection in a translation machine."

And yet I believe that we shall one day be able to achieve an automatic translation output which, even if it is not as good as one a good human translator is able to produce, may already be as good, or perhaps even better, than one a bad human translator would produce.

Papers on MT are nowadays heavily weighted on the side of the development of structural linguistic procedures for the solution of MT problems, and very rightly so. But many of these problems can be solved by lexicography. Dr. King has said here a few days ago that permanent memory devices are not necessarily as stupid as some may think. My paper will attempt to show this.

In my report "Current MT Research at the University of Washington" I mentioned that, in the course of lexicographical studies concerning the Russian-English MT-operational lexicon with whose elaboration we had been charged by the USAF, I soon realized that in certain types of cases of higher frequency it is possible to solve grammatical and non-grammatical problems by lexicography and lexicographical procedures alone--that is, without the necessity of logical procedures and logical machine operations. Since our sponsors had asked us to concentrate, at least during the initial phases of our project, on the elaboration of a bilingual lexicon, we decided to try to achieve an optimum of lexicography which would solve as many of our bilingual problems as possible. The results of this lexicographical work were published in our previous report [1] and supplied to our sponsors in the form of almost 170, 000 MT-operational Russian-English entries.

When I finished editing our first comprehensive report, I was convinced that we had actually achieved the optimum of lexicography we had intended. But hardly had that report been printed and distributed, when it occurred to me that I could vastly increase the number of instances in which an automatic system could be made to supply idiomatic translations without any logical operations and only on the basis of the information we are able to include in the bilingual lexicon. This new development I have already reported elsewhere [2] .

This unexpected development of a further extension of the area of MT problems that may be solved by lexicographical procedures alone encourages me to believe in the possibility that future research may uncover additional lexicographical procedures of this kind--with the pleasant result of increasing the degree of agreement of the MT product with the idiomatic requirements of the target language without necessarily increasing the number of logical machine operations. I should, therefore, like to outline in the following the results which have been achieved so far, hoping that this approach may also be taken up by others. The details will be found in the publications mentioned above.

Our MT research has, as is well known, been linked with the photoscopic memory device designed by Dr. Gilbert King. Since this permanent memory system has, or soon will have, a very large storage capacity, we were free to include in our bilingual lexicon any number of entries we thought necessary. In this lexicon we treated not only individual free and bound forms as lexical units, but also a number of uninterrupted idiomatic free-form sequences. It is necessary here to distinguish two kinds of idioms, namely:

1. Those that are idiomatic not only in the source language, but also in the target language, and therefore, should not be translated word-for-word. An example is the English "to fall in love with somebody" for which the idiomatic German equivalent is not mit jemandem *in Liebe fallen* but *sich in jemanden verlieben.*

2. Those which are completely shared by both languages and therefore can be translated word-for-word since such a translation does justice to the idiomatic requirements of the target language. An example is "he is a fox" corresponding to the German *er ist ein Fuchs*.

Shared idioms do not present any problems in MT. The unshared idioms, on the other hand, require special consideration. They should all be entered into the lexicon in toto, together with all their possible syntactic and paradigmatic transformations, and at their side should appear an idiomatic translation for every one of their transformations. Because of the limitations of the available equipment, we have until now been able to do this only in the case of the uninterrupted idiomatic sequences.

Another important distinction 1 had to make is the following:

1. Genuine Idioms like Chinese 費心 which literally translated means "to waste the heart", but which actually means "to put (somebody) to trouble", or "to thank (somebody)".

2. Pseudo-Idioms as, for example, such really non-idiomatic high-frequency expressions like "first of all" which, when translated word-for-word into German would be correctly understood although it would not be good German. The idiomatic German equivalent is erstens, in erster Linie or vor allen Dingen. Another example of this type, and moreover, one that represents an enormous number of expressions, is "the fundamental idea" which, when translated word-for-word into German, namely die grundlegende Idee is correctly understood and is even good German. The more idiomatic German equivalent would however be der Grundgedanke.

The word-for-word translation into German of "the fundamental idea" does not present any grammatical, semantic, or stylistic problem, and if English is the source language, then we can be satisfied with it. But if we are dealing with languages like Russian as the source language, word-for-word translations give us in most cases a translation product cluttered up with multiple grammatical alternatives reducing to a high degree the intelligibility of the output. Here the automatic supply of an idiomatic translation like German Grundgedanke would result in a substantial decrease of this superfluous clutter.

An important aspect of such expressions as "the fundamental idea" is that they are extremely numerous, represent high-frequency concepts, and that they are mostly not recorded in monolingual or bilingual dictionaries because everybody who knows the meanings of the free-form constituents of these expressions and is familiar with the rules of the languages, also understands the total meaning of the expression and does not need to look it up in a dictionary. The problem with which we are confronted here is the following. Since these expressions, which in most languages consist of more than one free form and represent high-frequency concepts, are not recorded, we would have to wade through an enormous corpus of publications to identify and collect them, unless there is a simpler way of obtaining a large number of them.

This simpler way I have described and exemplified in the paper [2] referred to above. Briefly, it is the following. The German language is well known for its great tendency and capability of forming compound words. There are two kinds of such compounds, those extemporized for the requirements of the moment and those of long standing which are recorded in both monolingual and bilingual dictionaries, and thus readily accessible. It is the latter which are important for us. They represent high-frequency concepts shared by all languages of importance for MT. If we utilize this conceptual experience available 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 important number of uninterrupted semantic units. In our MT lexicography, we can treat these as if they were idioms, and we shall obtain them with comparatively little effort--that is, without the necessity of an expensive and time-consuming search through large quantities of publications.

REFERENCES

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

[2] Erwin Reifler, MT Linguistics and MT Lexicography at the University of Washington, Proceedings of the International Conference for Standards on a Common Language for Machine Searching and Translation, Cleveland, Ohio, September 6-12, 1959. In the process of publication by Interscience Publishers, New York.