MACHINE LANGUAGE TRANSLATION

BY

ERWIN REIFLER

II. The Fundamental Problems

1. The Timeliness of MT

The sudden advent of MT development is no chance occurrence. It is the natural consequence of the coincidence of two human achievements, namely: a) a highly developed technology, and b) an advanced science of linguistics able to analyze and describe the bilingual or multilingual problems posed by this new venture of MT, and to formulate its requirements.

Ultimate success will depend on the cooperation of both. At this early stage, however, the primary task is linguistic. Once the linguistic problems have been solved, the engineers will know how to solve the engineering problems involved.

2. The Significance of MT

It is clear that the impact of MT on human culture and civilization will by far surpass that of the invention of book printing. Like the latter it will contribute immensely to the spread and exchange of knowledge. But book printing made knowledge more readily accessible only to large numbers of members of *the same speech community*; MT, on the other hand, will not know such limitation. On the contrary, the very purpose of MT is *the high-speed mass translation by machine from one language into one or more other languages*--that is, *the surmounting of the language barrier by automatic devices*--a purpose which transcends the interests of the individual speech community.

3. The Linguistic Prerequisites

MT presupposes that different languages can somehow be mechanically correlated. Such a correlation is, however, only possible if: a) they have something in common, and if b) what they have in common is expressed or expressible in distinctive signals.

This is actually the case. The speakers of all languages are endowed with the same kind of speech apparatus which they use in a more or less similar way. But also a large number of linguistic features is known to be either universal, widespread[1] or shared by at least two languages.[2] Certain common or comparable features concern the phonetic aspect of language.[3] Those, however, belonging to its logical aspect are especially numerous. They are found in the grammar[4] as well as in the lexicon.[5]

[1] Leonard Bloomfield, *Language*, pp. 270, 271. "A task for linguists of the future will be to compare the categories of different languages and see what features are universal or at least widespread...a form class comparable to our substantive expressions, with a class-meaning something like 'object,' seems to exist everywhere..."

[2] Joseph Vendryes,"La Comparaison En Linguistique," *Bulletin de la Société de Linguistique de Paris*, 1945, p. 5: "...àcertains égards l'anglai...pourrait admettre *une (classification), où il* figurerait par exemple à côté du Chinois; puisque, abstraction faite de toute parenté, évidemment hors de question, l'anglais et le chinois présentent, dit-on certaines ressemblances de structure."

[3] Vendryes, ibid., p. 7: "Assurément cette recherche d'une doctrine générale n'est pas nouvelle. Il y a une partie de la linguistique où elle est pratiquée depuis plusieurs décades et où elle a obtenu des résultats décisifs. C'est la phonétique..."

[4] Vendryes, ibid., p. 15: "...Une enquête générale sur les langues du monde fera connaître les lacunes et des manques dans l'usage des catégories grammaticales. Mais elle montrera aussi quelles sont les plus répandues, celles qui paraissent necessaires et qui répondent à un besoin de l'esprit humain." Cf. also footnote 1 above.

[5] Vendryes, ibid. pp. 16, 17: "...L'enquête qu'il y aurait lieu de faire sur l'évolution des faits grammaticaux pourrait naturellement être étendue aux faits de vocabulaire...La sémantique est une science qui doit aussi faire appel à la méthode comparative en l'appliquant à toutes les langues...On peut croire que les mêmes sentiments produisent à peu près les mêmes effets chez tous les peuples. Ces effets doivent se refléter dans le langage".

3.1. Semasiological Parallels

Such common or comparable features belonging to the logical aspect of language are due to independent parallel developments in the meanings of speech forms of languages which may be unrelated and distant from one another in space and time. These parallel developments are independent in the sense that they are linguistic coincidences although they betray the workings of a common human logic. A striking example of a widespread parallel is Chinese t'ung^{2} (童), meaning "child" as well as "pupil of the eye" (in the latter sense today written 瞳 --that is augmented by 目 which means "eye"). Many languages share this phenomenon of the association of the two, apparently incompatible, notions of "child" and "pupil of the eye" in one and the same word. English pupil (German Pupille) is itself derived from Latin pupilla meaning "little girl" as well (pupillus means "little boy") as also "pupil of the eye." The explanation for this phenomenon is the fact that whenever we look into somebody's eyes, we see there a small mirror image of ourselves.[6]

Another, but less common semasiological parallel is Chinese hsing^{2} (行) which means to go as well as also "it will do" like French aller and ca va and German gehen and es geht.

These are examples for parallels in the development of lexical or, better, non-grammatical meaning. An example for an independent parallel evolution of grammatical concepts is literary Chinese chih^{1}(之) which means to go as well as that, this. We see here the association of the concept of a demonstrative pronoun with that of a verb denoting "movement to another place." This phenomenon is paralleled in a number of languages, among them ancient Latin: both ire ("to go") and is, ea, id ("that, this") are traced back to an identical Indo-European ei--. The details I have discussed elsewhere.[7]

Another example is modern Mandarin che^{4}ke^{4} (這個)which, when stressed, is the demonstrative pronoun meaning "this." If, on the other hand, it is in an unstressed position, it represents the definite article. Analogically also i^{2}ko^{4} (一個), the Mandarin cardinal number meaning "one," corresponds to the English indefinite article "a, an" when in an unstressed position. This phenomenon of an evolution of the grammatical concepts of "definite article" and "indefinite article" from the earlier concept of a demonstrative pronoun or cardinal number, respectively, is paralled in many languages, among them Romance and Germanic languages (cf. the two uses of German der, die, das and of ein).[8]

3.2. The Ability of Language to Verbalize Concepts

Most important, however, for MT is the fact that "as to denotation, whatever can be said in one language can doubtless be said in any other...the difference will concern only the structure of the forms and their connotation."[9] Meanings and ideas expressed in one language can, for example, be expressed in another language by:

1. Idiomatic translation--that is a spoken or written form or a sequence of such forms peculiar to its linguistic system and imagery. An example is classical Greek ἐπιμελέομαι (with genitive), literally something like "I am making myself concerned in the direction of." To this corresponds classical Hebrew אני שם לבי אל, literally something like "I place my heart towards," and modern Chinese wo^{3} chu^{4} i^{4} (我 注 意), literally something like "I pour thoughts on" (followed by a direct object). The idiomatic translation of this concept into English is I pay attention to, and into German it is ich richte mein Augenmerk auf (or berücksichtige, bin aufmerksam auf, achte auf) which literally is something like "I direct my eyes' attention on to." Although we have here the different imageries of something like concern or care in Greek, "placing" one's heart in Hebrew, "pouring" one's thoughts in Chinese, "paying" attention in English and "directing" one's eyes in German, they all express very well the same notion.

2. Native expressions used metaphorically. An example is Chinese tien^{4} (電), "lightning," also used to translate English "electric, electricity," etc.

3. Native forms used in loan translations. For example "automobile" (auto meaning "self," mobile meaning "movable") to which modern Chinese tzu^{4} tung^{4} ch'ê^{1} (自 動 車) literally corresponds to (自 "self," 動 "move," 車 "carriage").

4. New spoken or written forms which are either:
   a) loanwords from another language, for example German Ersatz in English.
   b) new creations like Sanminismus to render Chinese San^{1} Min^{2} Chu^{3} I^{4} (三 民 主 義), the "Three People's Principles" of Sunyatsen.

[6] Cf. Erwin Reifler. "Linguistic Analysis, Meaning and Comparative Semantics," Lingua III, 4, Haarlem-Holland, August, 1953, pp. 381, 382, for semantic parallels for the same phenomenon in other languages.

[7] Erwin Reifler. "A Few Striking Examples Demonstrating the Contribution Comparative Semasiology Can Make Towards Historical Linguistics," Proceedings of the VIII. International Congress of Linguists, Oslo, August, 1957, pp. 622-25. In this paper I had pointed out that the archaic form of the Chinese character clearly indicated that the primary idea had been the verbal concept of "moving to another place" or "going thither"--that is, something like "to thither." This concept of "thithering" was then not only used dynamically as a verb, but also statically as "something thither,"--that is,a demonstrative pronoun corresponding in meaning to English yon or yonder.

[8] Cf. Erwin Reifler, "La Fission De L'Atome En Sinologie Á L'Aide De La Sémantique Comparative," Bulletin De L' Université L' Aurore, Shanghai, 1948, and the paper quoted in footnote 6 where other examples for such independent parallels in the evolution of non-grammatical and grammatical concepts will be found.

[9] Leonard Bloomfield, op. cit., p. 278.

4. The Linguistic Signals and Their Semantic Distinctiveness

The primary concern of all translation is meaning--the meaning intended by the original author-- and its translation from one linguistic system of signals into another linguistic system of signals. Consequently, the crucial problem for the success of MT is the *semantic* distinctiveness of the linguistic signals--that is, their ability to distinguish between different meanings, and the extent of this ability. The logical aspect of language, and any universal features discernible in the logical aspect, will be of prime importance for MT, whereas universals in phonetics can hardly have any bearing on it.

Furthermore, of the conventional signals of language only the phonic and graphic signals[10] fall within the sphere of interest of MT as initiatory stimuli for a correlation between languages by machine. There are, however, three reasons why comparatively little is being done at present in MT for speech ("phonic MT") and why most efforts are being directed towards the development of MT for written texts (graphic MT). These reasons are:

a) A number of engineering problems have still to be solved to permit MT based on the speech of *any* member of a speech community.

b) The conventional phonic signals are in a number of important languages often less distinctive than their corresponding written forms.[11]

c) It is easier for writers than for speakers to adhere closely to the conventions of their language, and they are more likely to do so.

MT research is, therefore, at present largely concerned with the conventional *graphic* forms of the linguistic signals of a language although also these are often not distinctive. This distinctiveness is the greater the larger the meaningful unit considered, and it decreases with the size of the latter. It is greatest in idiomatic expressions consisting of more than one free form, and this fact enables us, as we shall see later, to use purely lexicographical procedures and yet achieve an idiomatic MT of such idioms which no human translator could do better. The distinctiveness is, of course, smallest in the case of meaningful units consisting of only one free form and considered in isolation. It is worth while at this point to try to ascertain the size of the problem MT faces here with regard to the phonic and graphic distinctiveness of minimum free forms. If we denote origin by O, meaning by M, pronunciation by P, and spelling by S, and, furthermore, indicate agreement and disagreement between two words of a language in origin, meaning, pronunciation and spelling by prefixing a plus or a minus sign, then there are the following *theoretically possible* permutational combinations:

1.	+O	+M	+P	+S
2.	-O	+M	+P	+S
3.	-O	-M	+P	+S
4.	-O	-M	-P	+S
5.	-O	-M	-P	-S
6.	+O	-M	-P	-S
7.	+O	+M	-P	-S
8.	+O	+M	+P	-S
9.	+O	+M	-P	+S
10.	+O	-M	+P	+S
11.	+O	-M	-P	+S
12.	-O	+M	+P	-S
13.	+O	-M	+P	-S
14.	-O	+M	-P	+S
15.	-O	+M	-P	-S
16.	-O	-M	+P	-S

[10] i.e., the meaningful constituents of speech and their written form.

[11] So, for example, in Chinese and Japanese, but also in languages like English and French--that is, in languages with a historical form of writing. Striking examples are the three English words *so, sew, sow* which have identical pronunciations.

Of these the first is meaningless, the second unlikely. No problems for either phonic or graphic MT are presented by:

No. 5, i.e. the two entirely different words good and evil.
No. 6, i.e. the two cognate words cattle and chattel.
No. 7, i.e. the two cognates of identical meaning yon and yonder.
No. 8, i.e. the graphic variants draught and draft, gaol and jail, licence and license.
No. 12, if there are such.
No. 14, if there are such.
No. 15, the case of the perfect synonyms, if there are such.

The following two cases are of importance only to phonic MT:

No. 13, i.e. the cognate pairs to and too, holy and wholly.
No. 16, i.e. the unrelated pairs no and know, die and dye, to (or too) and two.

This leaves only the possibilities Nos. 3, 4, 9, 10, and 11. Of these the following two are of importance to both phonic and graphic MT:

No. 3, i.e. seal (German "Siegel") and seal (German "Seehund"),
lie (German "liegen") and lie (German "lügen"),
die (German "Würfel") and die (German "sterben"),
rose (German "Rose") and rose (German "erhob sich"),
leaves (German "Blätter") and leaves (German "verlässt"),
last (German "letzter") and last (German "dauern")
No. 10, i. e. sheep (German "Schaf") and sheep (German "Schafe"),
love (German "Liebe") and love (German "lieben"),
fell (German "fällen") and fell (German "fiel").

The remaining cases are of importance only for graphic MT. They are:

No. 4, i.e. lead (German "Blei") and lead (German "führen").
No. 9, i.e. cónvert (German "Bekehter") and convért (German "bekehren").
No. 11, i.e. read (German "lesen") and read (German "las, gelesen").

Since, as I have already pointed out above, MT research is at present concentrating on graphic MT, it is not concerned with homophones[12] but with homographs.[13] Consequently the problems exemplified in Nos. 3, 4, 9, 10, and 11 are of greatest importance for it.

In the preceding tabulation I have included the criterion of origin because it brings out a detail which may be of some consequence in MT development. It shows that homographs may either represent related speech forms as those in numbers 9, 10, and 11, or unrelated ones as those in numbers 3 and 4. The two speech forms represented by read in number 11 are, of course, members of the same verb paradigm. But we could, in our MT-lexicography, also treat homographic cognates of the high-frequency type "substantive/verb," like love in number 10, or convert in number 9, as if they were members of a "super-paradigm," and then devise a logical routine for the automatic pinpointing of the intended grammatical meaning (either substantive or verb) for this whole type which in English has a large membership and is still very productive. Such a procedure is hardly advisable in the case of non-cognate homographic speech forms like those exemplified by lie, die, rose, leaves, last in number 3, and lead in number 4, which in English are neither numerous nor productive.

If we ignore the criterion of origin--as we well may in a purely descriptive survey of phono-semantic and graphio-semantic distinctiveness[14] --then the types numbers 1 and 2, 3 and 10, 4 and 11, 5 and 6, 7 and 15, 8 and 12, 9 and 14, 13 and 16 coalesce, giving us eight theoretically possible patterns of which only three are of practical significance for graphic MT. These are the patterns labeled by the numbers 3/10, 4/11, and 9/14, namely all those whose last criterion is +S--that is homography.

5. Conventional Means of Increasing the Semantic Distinctiveness of Individual Free Forms

Messages often consist not only of primary meaningful units such as sentences, clauses, idiomatic expressions, phrases and even individual free forms, but also of certain concomitant features which contribute to the total meaning. Speakers as well as writers often use such additional means to achieve semantic distinctiveness of ambiguous individual free forms even when the narrow or wider linguistic environment would sufficiently pinpoint the intended meaning. In listening to speakers we are aware of their tacit utilization of situations, and accompanying actions such as "facial expressions, mimicry, tone of voice..., insignificant handling of objects

[12] i.e. different words with identical pronunciations, as those mentioned in the preceding footnote.

[13] i.e. different words with identical spellings.

[14] i.e. the extent to which different meanings are distinguished by differences in pronunciation or writing, respectively.

..., and, above all, gesture."[15] In writing we are familiar with the phenomenon of conventional supplementary symbols, employed as differentiating features already in ancient times. Examples are in the case of alphabetic scripts, the distinction between capital and small letters and the dots of the Umlaut in German, the spiritus symbols in Greek, the accents in Greek and in other languages, the vowel points, the dagesh and the shwa in Hebrew. Examples for such conventional differentiators in non-alphabetic scripts are significs in Egyptian hieroglyphic writing and in Chinese (the so-called "radicals"), the tone marks in Chinese, the Japanese kana symbols at the side of Chinese characters, or also, as we shall see below, Chinese characters at the side of Japanese kanas.

It is doubtful whether phonic MT will ever be able to utilize secondary features of speech to reduce the size of the multiple meaning problem of homophones. On the other hand, graphic MT will make full use of the supplementary graphic symbols. The latter are, however, not always obligatory. German Umlaut dots (or their substitute, the letter e after the vowels a, o, and u) are always used, but there are German texts in which all words are written with small initials so that the non-grammatical and grammatical distinction between the capital and small initials of individual free forms in non-first positions[16] is absent. The Hebrew vowel points and the Chinese tone marks are mostly omitted. The absence of such supplementary symbols increases the number of alternative semantic possibilities. But there are also many cases in which the lack of graphio-semantic distinctiveness is not due to the absence of supplementary symbols. For all cases of multiple meaning it is necessary to devise ways and means to enable an automatic system to pinpoint the intended meaning in consideration of the narrow or wider environment of ambiguous forms. These ways and means will be outlined further below.

6. Different Writing Systems

MT development has to deal with different writing systems which are either alphabetic, syllabic,[17] ideographic,[18] a combination of the alphabetic and ideographic[19] or of the syllabic and ideographic[20] systems. In the last two cases the ideographs are either the primary or the supplementary symbols. Research is at present being carried on in the United States, in Japan, and in other countries for the purpose of developing electronic readers which will read and code printed source language material and feed it into the automatic translation system for further processing.

The following tables exemplify the different writing systems MT development has to consider, and demonstrate the use and significance of supplementary symbols which are often omitted from printed texts. It is helpful to distinguish here several cases, depending on the type of multiple meaning involved, and on whether or not differences in meaning are associated with differences in pronunciation.

6.1. Conventional Supplementary Symbols Indicative of Semantic Differences Associated with Phonic Differences

1. Multiple Non-Grammatical Meaning

Greek	ιέναι	which is:			
	either	ἰέναι	"to go," or	ἱέναι	, "to send"[21] (both verbs)
Hebrew	מלך	which is:			
	either	מֶלֶךְ	"king," or	מֹלֶךְ	, "Moloch"[22] (both substantives)
Chinese	蹶	which is:			

[15] Leonard Bloomfield, op. cit., p. 39.

[16] The difference between Los (lot) and los (loose) is both non-grammatical (the lexical meaning "lot" versus the lexical meaning "loose") and grammatical (Los is a substantive, los a predicative adjective). The initials of the following forms are conventionally capitalized:

a) All forms of pronouns used in address instead of du, and, in letter writing, all pronouns (including du) referring to the addressed person.
b) All adjectives derived from personal names by the suffix -isch.
c) All adjectives, pronouns and ordinal numbers in titles and in historical and geographical names.
d) All invariable word forms with the suffix -er, derived from place names, and the names of provinces or federal states.
e) All substantives with the exception of certain petrified forms and certain forms used in idiomatic expressions.

[17] For example, the Ethiopic system.

[18] For example, the modern Chinese system.

[19] In Korean.

[20] In Japanese.

[21] The different spiritus indicate both different pronunciation (for ancient Greek) and meaning.

[22] The difference in vowel pointing distinguishes both the different pronunciations and meanings.

either ꜀蹶 , "to stumble," or ᶜ蹶 , "to kick," or 蹶ˀ "to excite"[23] (all verbs)

Japanese 里 which is:

either 里サト , "native place," or 里リ , "Japanese mile"[24] (both substantives)

(2) Multiple Grammatical Meaning

Greek πoυ which is:

either ποῦ , "where," or πού , "anywhere"[25] (interrogative versus indefinite adverb)

Hebrew מלך which is:

either מֶלֶךְ , "king," or מָלַךְ , "he ruled"[22] (substantive versus verb)

Chinese 種 which is:

either ᶜ種 , "seed," or 種ˀ , "to sow"[23] (substantive versus verb)

Japanese 乱 which is:

either 乱ミダレ , "disorder," or 乱ミダル "to throw into disorder"[24] (substantive versus verb)

(3) Multiple Non-Grammatical and Grammatical Meaning

Greek εἰμι which is:

either εἶμι . "I shall go," or εἰμί , "I am"[25] (different lexical meaning and tense)

Hebrew אלה which is:

either אֵלָה , "oak," or אָלָה , "he cursed," or אֵלֶּה , "these"[22] (different lexical meaning and form class)

Chinese 惡 which is:

either 惡ˀ , "to dislike," or ꜀惡 , "where? why? how?"[23] (different lexical meaning and form class)

Japanese 生 which is:

either 生ウム , "to beget," or 生ナマ , "raw," or 生セイ , "life"[24] (different lexical meaning and form class)

6.2. Conventional Supplementary Symbols Indicative of Semantic Differences Not Associated with Phonic Differences

(1) Multiple Non-Grammatical Meaning

Japanese ヒ which is:

either 日ヒ , "sun," or 火ヒ , "fire"[26] (both substantives)

Japanese ウム which is:

either 熟ウム , "to ripen," or 生ウム , "to beget"[26] (both verbs)

(2) Multiple Grammatical Meaning

Japanese 多 which is:

[23] The different positions of the tone mark distinguish between different pronunciations and meanings.

[24] The different kana symbols on the right of the same Chinese character indicate the different pronunciations and meanings.

[25] The different accents indicate both different pronunciations (for ancient Greek) and meaning.

[26] Here the Japanese kana remains unchanged whereas the Chinese characters are the supplementary differentiating symbols.

either 自カラ , "since" (the preposition) or 故カラ, "since" (the conjunction)[26]

German 'band' which is:

either band, "he bound," or Band, "ribbon, volume"[27] (verb and substantive)

(3) Multiple Non-Grammatical and Grammatical Meaning

Japanese キ which is:

either 木キ , "tree," or 黄キ , "yellow"[26] (different lexical meaning and form class)

7. The Concurrent Pinpointing of Intended Non-Grammatical Meaning by the Determination of Grammatical Meaning

The preceding tables include examples showing that supplementary symbols sometimes simultaneously pinpoint both non-grammatical and grammatical meaning. I have, however, already pointed out before that such supplementary symbols are often not used. There are, moreover, many instances in which either no supplementary symbolization is available or where it is not effective. Examples for the latter case are German substantives when they introduce sentences. Since not only substantives but every word at the beginning of a German sentence has an initial capital letter, the form class of a substantive in such a position cannot be recognized by its initial letter. A striking example for the ambiguity that may be caused by this loss of effectiveness of German initial capital letters when they are in sentence-first positions is Dichter ist der Hahn geworden. This sentence may occur in a versified fable and the intended translation could then well be "A poet has the cock become." Otherwise the correct translation would be "Tighter has the faucet become."

In such cases it is necessary that the MT system determine both types of meaning, the non-grammatical and the grammatical, by a consideration of the narrow or wider context. It will, of course, first try to determine the intended grammatical meaning on the basis of the structural analysis of the linguistic environment of the ambiguous form. Once the grammatical meaning has been determined, the automatic system will try to determine also the intended non-grammatical meaning in consideration of the grammatical meaning. Frequently, however, this second step will not be necessary, because in many instances the pinpointing of the intended grammatical meaning simultaneously pinpoints the intended non-grammatical meaning. The following Chinese example will illustrate this:

The speech form represented by 光 (kuang^1) is used as a substantive, an adjective and an adverb. As a substantive it means "light;" as an adjective it corresponds either to English "light" (as opposed to dark), "shining," "bald," or "bare." As an adverb it means "only." The semantic relationship between these meanings is clear, except in the case of "only." But even the position of "only" in the scale of the meanings of 光 becomes clear if we compare the German semasiological parallel "bloss" which means "bare" as well as "only": "er hat bloss zwei Dollar" originally meant "he has barely two dollars," but it has come to mean "he has only two dollars," just like modern Mandarin 他光有兩塊錢 / The automatic translation system would, therefore, in the case of 光 first have to determine the form class--that is, whether it is a substantive, an adjective or an adverb, because the determination of the form class would here be tantamount to the pinpointing of the non-grammatical meaning.

In the case of the adjectival use of 光 this procedure is not quite satisfactory because it is not able to make a decision between the four alternatives "light/shining/bald/bare." If no procedure were available which could make such a decision, the automatic system would on the output side have to supply this whole string of alternatives, and it would have to be left to the human reader of the MT product to select that alternative which best fits the context. There is, however, a procedure which permits the automatic system to supply idiomatically correct translations in such cases. This procedure is based on the following facts. Of the meanings of adjectival 光 only one, namely "shining" may qualify a large variety of substantives, whereas in each of the meanings of "light," "bald," and "bare" it occurs as a qualifier of only a comparatively limited number of substantives.[28] Examples are:

光射	(literally: shining shot) "light beam, ray."
光景	(literally: shining or visible scape) "landscape, aspect of things, circumstances."
光頭	(literally: shining or visible head) "bald head."
光身	(literally: shining or visible body) "bare body."

Now we can treat the comparatively small number of these adjective-substantive phrases as idiomatic ex-

[27] Here the difference between small and capital initial is significant.

[28] In order to simplify the presentation, we ignore here the predicative use of 光

pressions,[29] and code each as a semantic unit into the permanent memory device of the automatic translation system. To each we add an idiomatic target language translation. If we do that, then the automatic translation system will in such cases always supply idiomatic translations.[30]

8. The Problem of Idioms

For the purposes of MT we have to distinguish between two kinds of idioms, namely:

a) *Monolingual Idioms*--that is, idioms of the source language that are not shared by the target language. An example is German *das geht nicht*, literally "that goes not," but meaning "that cannot be done."
b) *Bilingual Idioms*--that is, those that are shared by both the source and the target language. German *das geht nicht* is such a bilingual idiom if either French or modern Mandarin is the target language (cf. "çela ne va pas" and 這個不行).

The problem of idioms has in MT development always to be considered in the light of the two languages concerned in the translation process--that is, in the light of "source-target semantics." Bilingual idioms permit a word-for-word translation. This is, however, not possible in the case of monolingual idioms which require a free translation that conforms to the idiomatic requirements of the target language. In order to enable an automatic translation system to supply target-idiomatic translations for monolingual idioms we have to use the procedure already mentioned in the preceding section: we have to include the source language idiom *in toto* in the automatic bilingual lexicon, together with the idiomatic target language translation.

We have, furthermore, to distinguish between paradigmatic idioms, namely those containing semantic units which are members of a paradigm and therefore may appear in different paradigmatic forms, and those that are not paradigmatic. German *das geht nicht* is paradigmatic because it may also appear in the forms *das ging nicht* (that could not be done), *das ist nicht gegangen* (it has not been possible to do that), *das wird nicht gehen* it will not be possible to do that), etc., etc., or in the different word order required in dependent clauses, such as *dass das nicht geht* (that that can not be done). An example for a non-paradigmatic idiom is the English idiomatic expression *to make a long story short* in which the verb *make* is never conjugated nor the substantive *story* ever declined. If a monolingual idiom is paradigmatic, it has to be entered into the automatic permanent memory with all the permissible transformations due to the inflexion of forms and/or the requirements of word order, and together with all the idiomatic translations for these transformations.[30]

Another distinction which is proving very useful in MT is that between "genuine idioms" and "pseudo-idioms." German *das geht nicht* is a genuine idiom in terms of the semantic peculiarities of the English language because a literal translation, namely something like "that goes not," does not make sense. A literal translation of German *in erster Linie*, namely something like "in first line," not only does make sense, it even makes good sense and will, moreover, be correctly understood in the translated context. It is, therefore, not a genuine source-target semantic idiom. "In the first line" is, however, not good English: the idiomatic translation is "in the first place" or "first of all." It we want the automatic translation system to supply idiomatic translations also in such cases, we have to treat them *as if they were idioms* and enter them *in toto* into the automatic permanent memory together with the idiomatic translation. Such a procedure is, of course, only possible in the case of high-frequency expressions which, since they have a high-frequency rating, are not very numerous. German *in erster Linie*, English first of all, are good examples for such high frequency expressions.

It has to be emphasized here that idioms, whether they are genuine or pseudo-idioms, are graphio-semantically highly distinctive. In the English monolingual idiom[31] *to bark up the wrong tree*, for example, a form of the verb *to bark*, the adverb *up*, a form of the substantive *tree* and the article *the* have to co-occur to make it an idiom. It is this graphio-semantic distinctiveness, which makes it possible to use the procedure outlined above, and will permit future translation machines to supply idiomatic translations for idioms no human translator could do better. Thus idioms will never be an obstacle to MT; it is rather the non-idiomatic portions of the source-language text which present the most difficult problems for the automation of the translation process.[32]

9. The Problem of Unpredictable Forms

When discussing the *ability of language to verbalize concepts*, I mentioned and exemplified a number of

[29] And they certainly are idioms in terms of, for example, Chinese-English "source-target semantics" because otherwise the literal translation given for them above in parentheses would be satisfactory.

[30] Permanent automatic memories with a large storage capacity and a very low access time are already being developed in the United States. They will make the procedure outlined above entirely feasible. I may add that it is relatively simple to insure that an automatic system recognize idiomatic sequences, for it need only be able to identify the longest source text constituent for which the permanent memory device has an equivalent entry. This procedure is one of the features of the scheme of the University of Washington MT Project.

[31] That is, monolingual in consideration of the languages known to the present writer.

[32] Cf. Erwin Reifler, "The Machine Translation Project at the University of Washington, Outline of the Project, §3.3: Meaning," *Linguistic and Engineering Studies in the Automatic Translation of Scientific Russian into English, Technical Report*. Seattle: The University of Washington Press, 1958.

means which are at the disposal of a language and enable it to express whatever can be expressed in any other language. As the last means I mentioned the use of new forms which are either borrowed from another language (loanwords) or new creations. Among these new forms we have to distinguish between the following two groups:

1. Those which at the time of the compilation of a MT lexicon are already recognized as members of the vocabulary of the source language and, therefore, can be considered for inclusion.
2. Those which are not yet known to the compilers of the MT lexicon and, therefore, can not be considered for inclusion.

Whenever a new form not included among the entries of the permanent MT memory is fed into the translation system, it will not be identified and can, therefore, not be translated. The MT system can, however, be so designed that it transfers every unidentified input form to the output either in its original script form, or in a transcription which is either in a special code or in the writing system of the target language. The MT system may, moreover, put out such a transcribed source language form in a distinguishing colour[33] so that the MT researcher who studies the translation output can easily detect and collect all semantic units of the source language which are not yet included in the automatic lexicon.

There is, however, one type of source language forms which an automatic translation system can be made to translate correctly, even though they are not yet known at the time of the compilation of the MT lexicon and not included in it. This is the ever-growing number of extemporized, and thus unpredictable, compounds which in many languages, foremost among them German, are created anew by speakers and writers for the requirements of the moment. Such compounds are entirely made up of meaningful bound forms which occur also as well-known minimum free forms and which, therefore, can all be included in the automatic memory. An example for such an extemporized compound is German Marssprachenforschungsgesellschaftsbericht (literally "Mars Languages Research Society Report") which I have just made up to demonstrate the problem. Mars, Sprachen, Forschung(s), Gesellschaft(s), and Bericht occur all as free forms and will any way form part of the MT lexicon. The automatic MT system can be so designed that, whenever such a new compound is fed into it and not identified in the permanent memory, it will be dissected into the constituent forms which have free-form occurrence, are consequently found in the permanent memory and can therefore be translated by the system.[33]

This means that the automatic translation system will actually not translate the compound form, but rather its constituents. In the very numerous cases exemplified by Marssprachenforschungsgesellschaftsbericht this will amount to the same, but there are compounds which require special machine procedures in order to make sure that only the correct dissections, identifications and translations are made. A striking example is German Dichterinbrunst. The correct dissection is Dichter/inbrunst--that is, "A poet's fervour (or ardour)." Both Dichter (poet or poets) and Inbrunst (fervour, ardour) will be included in the permanent MT memory, but so will also dicht (tight, dense), in (in), Dichterin (poetess) and Brunst (sexual desire). If the translation system makes, for example, the wrong dissection Dichterin/brunst, it will supply the wrong translation," sexual desire of a poetess."

In my researches conducted in the summer of 1952 under a grant from the Rockefeller Foundation I found that two arrangements have to be made in order to ensure that the automatic MT system will in such cases always make the right dissection. One is the provision that the matching mechanism of the translation system always identifies first the longest constituent for which the permanent memory contains an equivalent.[30] The matching mechanism will, under this provision, first identify the longest lefthand constituent Dichterin--if right-to-left matching is used, and the longest right hand constituent--inbrunst if left-to-right matching is used.

The second provision is concerned with what I have called the "X-factor" problem in compounds. An "X-factor" is a constituent of a compound which may itself occur as a free form, or form a semantic unit either with the preceding or the following constituent of the compound. In Dichterinbrunst, -in- is such an "X-factor." The solution to this problem takes into account certain linguistic principles at the basis of the creation of compounds. As far as German -in- is concerned, the rule says that feminine substantives ending in "-in" can not be lefthand constituents of compound substantives. If this rule is included in the program of the automatic MT system, the latter will in such cases discard the result of right-to-left matching and retain only the result of left-to-right matching, thus arriving at the correct dissection Dichter/inbrunst. For the detailed discussion, exemplification and logical procedure the reader is referred to the special report on this problem.[34]

10. The Problem of Multiple Grammatical and Non-Grammatical Meaning

In the discussion and exemplification of the semantic distinctiveness of the linguistic signals, and in the tables exemplifying the supplementary symbols of the different writing systems I have already had occasion to distinguish between multiple grammatical and multiple non-grammatical meaning. The ambiguous script forms sheep (either singular or plural), love (either substantive or verb), fell (either the present tense of the verb to fell or the past tense of the verb to fall), convert (either the substantive ćonvert or the verb to convért) and read (either the present tense or the past tense of the same verb) are examples for multiple

[33] This is actually the procedure followed in the University of Washington Machine Translation Project.

[34] Erwin Reifler, "Mechanical Determination of the Constituents of German Substantive Compounds," MT, Vol. 2, No. 1, pp. 3-14. M. I. T., July, 1955, and "The Mechanical Determination of Meaning," Machine Translation of Languages, William N. Locke and A. Donald Booth, The Technology Press of M. I. T., and John Wiley & Sons, Inc., New York, 1955, pp. 144-148.

grammatical meaning. The ambiguous script forms *seal* (either German "Siegel" or "Seehund," in either case a substantive) and *lie* (either German "liegen" or "lügen," in either case a present tense verb) are examples for multiple *non-grammatical* meaning. On the other hand, the ambiguous script forms *die* (either a substantive corresponding to German "Würfel," or a verb corresponding to German "sterben"), *rose* (either a substantive equivalent to German "Rose," or a past tense verb equivalent to German "erhob sich"), *leaves* (either a plural substantive meaning German "Blätter," or a present tense, singular, third person verb meaning German "verlässt") and *last* (either an adjective translating German "letzter," or a verb translating German "dauern") are each simultaneous examples for both *multiple grammatical and multiple non-grammatical* meaning.

These have all been examples for grammatical and/or non-grammatical ambiguity due to graphio-morphologic identity. The ambiguity of one minimum free form may not affect the interpretation of other minimum-free-form constituents of a sentence. An example is the well-known Latin grammar school joke *qui patrem suum necat non peccat* which plays upon the double, both grammatical and non-grammatical, meaning of *suum* (either "his" or "of the swine") and, therefore, may be translated either "who kills his father does not sin" or "who kills the father of swine does not sin." The alternative translations "of swine" or "his" do not affect the translation of *qui patrem...necat non peccat* which remains "who kills...father...does not sin." In other cases the interpretation of other minimum-free-form constituents of the sentence may be seriously affected, as, for example, in the previously quoted German verse line *Dichter ist der Hahn geworden* (either "tighter has the faucet become" or "a poet has the cock become"). If *Dichter* is an adjective and means "tighter," then *Hahn* must be translated "faucet," but if it is a substantive and means "poet," then *Hahn* must be translated "cock." In such cases MT will have to consider the wider context in order to determine whether a faucet or a cock is intended.

Syntactic identity is also frequently a cause of ambiguity. If, for example, the German subject would always precede, the direct object always follow the verb, then no wider context would be needed to determine the meaning of such sentences as *die Löwin biss die Schlange* or *das Ungeheuer verschlang das Wasser*, and the task of MT would here be much simpler. But without a consideration of wider context the first may mean either "the lioness bit the serpent," or "the serpent bit the lioness," the second either "the monster swallowed the water," or "the water swallowed the monster."

Apart from the syntacto-hierarchic relationships within the principal and the dependent clause, MT research has also to investigate those between dependent clauses, and between such clauses and the principal clause. Most MT projects are at present concerned with such studies for the purpose of elaborating a hierarchy of operational rules for the automatic resolution of all grammatical and non-grammatical source-target problems, or, if we look at the task from the point of view of the desired translation product, for the automatic selection of target alternatives which represent the meaning or meanings intended by the original author.

Multiple *grammatical* meaning presents the smaller problem of the two. Grammatical ambiguity is frequent in isolation, *but* rare in a linguistic environment. A thorough knowledge of the operational principles and rules concerning the morphologic and syntactic behaviour of a language will, it is now generally felt, enable the MT linguist to elaborate logical procedures on the basis of which the MT engineer will be able to develop machine programs for the automatic pinpointing of intended grammatical meaning. There are probably many different ways in which an automatic system can be made to extract from a foreign text all relevant grammatical information. One such scheme, a simple procedure for the automatic determination of the form classes of the minimum-free-form constituents of an input text, I have outlined in a previous publication.[35]

Multiple non-grammatical meaning presents the most formidable problem in MT. In the case of genuine or pseudo-idioms--that is, source-target idiomatic free-form sequences of the input text--lexicographical procedures alone will be sufficient to ensure an idiomatic translation because such sequences are, as I have already emphasized earlier, graphio-semantically highly distinctive. In all other cases of multiple non-grammatical meaning, the prerequisite for the reduction of ambiguity will be the structural analysis of the source language text by the automatic system and the solution of all *grammatical* ambiguities of the text section in question. The determination of the first coincides in some cases, as we have seen earlier, with the determination of the second. In other cases it will be necessary for the automatic MT system to scan the syntactically connected narrow or wider context for clues that may pinpoint the intended non-grammatical meaning.

It is at present still doubtful whether it will ever be possible to resolve all problems of MT ambiguity by machine. There are those who believe that the non-grammatical meaning aspect presents complexities of an astronomical order which makes it advisable to forego here attempts at mechanization. Others, on the other hand, believe that the limitation to scientific publications, and the further limitation to a single branch or sub-branch of science, will sufficiently reduce the semantic problem to make complete mechanization of the translation process feasible. Consequently they reject the creation of large general MT lexicons and propose the creation of comparatively small glossaries, the so-called *idio-glossaries*, containing the highly specialized and, therefore, limited vocabulary of idiolects--that is, sections of a language characteristic of a branch or sub-branch of human knowledge.

The present writer does not believe that such a limitation is necessary. Such a limitation would only reduce the multiple non-grammatical meaning problem of the scientific and technological vocabulary, but would not solve that of the general language vocabulary which all branches of science and technology share. The non-grammatical meaning problem of the scientific and technological vocabulary, on the other hand, does not present any serious difficulties. On the contrary, our researches at the University of Washington Machine Translation Project show that this problem is amenable to an automatic solution.[36]

[35] Erwin Reifler, "The Mechanical Determination of Meaning," op. cit., pp. 161-63.

[36] Erwin Reifler, "The Machine Translation Project at the University of Washington, Outline of the Project," §2.2, op. cit.

The present writer believes that it is still too early to make safe predictions concerning the extent to which it will be ultimately possible to mechanize the translation process. Whatever the ultimate degree of success, also incomplete solutions will prove beneficial because they will relieve at least part of the burden of the human translator. Some present preliminary results may already be useful for some purposes.[37] It is in any case worth while trying to find out whether there are any insuperable natural boundaries preventing full success in this new adventure of the human mind.

11. Samples of Predicted Russian-English MT Output Based on Preliminary Research Results at the University of Washington MT Project[38]

In order to enable the reader to form an opinion about the type of MT already attainable today, we include in the Appendix to Part One "simulated machine translations" elaborated by hand on the basis of the Russian-English "MT-operational lexicon" prepared by the staff of the University of Washington Machine Translation Project. The Russian originals have been published in our previous report.

The reader of the simulated MT output is asked to keep in mind that such a "translation" is by no means to be taken as the final goal of MT development, although it may, as I have pointed out above, already be useful for some purposes. We consider it only as an intermediate result, to be studied in order to determine to what extent we have already succeeded, what has still to be done and what should be undertaken next.

12. Conclusion

Compared with other modern achievements, the development of such a highly complicated matter as mechanical translation seems to be proceeding at a surprising speed. Although it is today mainly a linguistic problem, let us not forget that this is so only because of the incredibly rapid pace at which our technology has progressed for some time. Without the previous development of electronic computing devices nobody would have given much thought to the possibility of mechanizing the translation process. It is, I believe, only right to acknowledge here that it is the natural sciences and the engineer who in the last analysis are making mechanical translation possible.

But how will this new threatening expansion of the empire of THE MACHINE affect our lives? Will it make superfluous the study of foreign languages and thus take away the bread of language teachers and translators? Will it lower the literary level of translations, or, worse still, will it have a corrupting influence on our natural languages and the artistic form of works written in them? What horrible consequences may all this have for human culture? Will these translation machines be the beginning of worse to come? Will they be the forerunners of teaching machines and learning machines taking the place of both teachers and students with the result that factories will ultimately replace universities? I know from many conversations, and these not with people outside the teaching profession, but with teachers on the university level, that such nightmarish questions are influencing the attitude of many with regard to the coming of translation machines.

I believe that just the opposite will be the case. This new venture requires the elaboration of so many linguistic details; the number of languages, which are or will become important because of their literary products, is so large that many more language experts will be required for this truly gigantic task. Thus the development of mechanical translation will rather encourage the teaching and learning of foreign languages. The human translator, on the other hand, will not only not suffer, he will actually benefit. He will use translation machines for the high-speed rough translation in bulk. Thus he will be able to devote all his working time to the polishing of the crude translation products supplied to him by the machine in an unbelievably short time. But he will need to do this only if publication is desired, and only as far as literary merits are involved. Consequently, he will be able to increase enormously the number of works he wants to translate during his life span. Nor do I believe that universities will disappear. On the contrary, they will become much more productive and efficient because of the quicker and wider accessibility of an incomparably larger number of important works of foreign origin.

But however one may feel about the progressive substitution of mechanical operations for the work of human beings, there can be no question about the growing need for mechanical translation. The ever-increasing volume of important publications in a multiplicity of languages, the insufficient number of competent translators and the time consumed in translation all justify the search for a mechanical solution to the problem of high-speed mass translation.

[37] Cf. the following section K.

[38] Section 11 differs from the original in the volume, Digitale Informationswandler, published by Friedrich vieweg & Sohn, Braunschweig, Germany, since it had to be adjusted to the requirements of the present report.