[*Mechanical Translation*, vol.3, no.3, December 1956; pp. 81-88]

The Linguistic Basis of a Mechanical Thesaurus †

M. A. K. Halliday, Cambridge Language Research Unit, Cambridge, England

The grammar and lexis of a language exhibit a high degree of internal determination, affecting all utterances whether or not these are translated from another language. This may be exploited in a mechanical translation program in order to cope with the lack of translation equivalence between categories of different languages, by the ordering of elements into systems within which determination operates and the working out by descriptive linguistic methods of the criteria governing the choice among the elements ranged as terms in one system. Lexical items so ordered form a thesaurus, and the thesaurus series is the lexical analogue of the grammatical paradigm.

A FUNDAMENTAL problem of mechanical translation, arising at the levels of both grammar and lexis, is that of the carry-over of elements ranged as terms in particular systems; i.e., systems established non-comparatively, as valid for the synchronic and syntopic description of what is regarded for the purpose as 'one' language. The translation process presupposes an analysis, generally unformulated in the case of human translation, of the source and target languages; and it is a commonplace that a one-to-one translation equivalence of categories - including not only terms within systems but even the systems themselves - does not by itself result in anything which on contextual criteria could be called translation. One might, for example, be tempted to give the same name 'aspect' to two systems set up in the description respectively of Chinese and English, on the grounds that both systems are the grammatical reflection of contextually specified categories of a non-absolute time-scale in which components of a situation are ordered in relation to one another; not only would the terms in the systems (e.g. Chinese and English 'perfective') not be translationally identifiable: not even the systems as a whole (unless a neutral term was introduced to universalize them) could be assigned translation equivalence.

---

† This is one of a series of four papers presented by the Cambridge Language Research Unit to the October 1956 Conference on Me-, chanical Translation (for abstracts see MT, Vol. II, No. 2, pp. 36-37).

Syntax

Where translation is handled as a function between two given languages, this problem can be met by a comparative description of the kind that has come to be known as 'transfer grammar', in which the two languages are described in mutually (or unilaterally) approximating comparative terms. For mechanical translation this is obviously unsatisfactory, since each language would have to be analyzed in a different way for every new language on the other end of the translation axis. On the other hand the search for categories with universal translation validity, or even with validity over a given limited group of languages, whether it is undertaken from within or from outside language, could occupy many years; and while the statistical survey required for the intralinguistic approach would be, for the linguist, perhaps the most pleasing form of electronic activity, the pursuit of mechanical translation cannot await its results!

In practice, therefore, we compromise, and make a descriptive analysis of each language which is at the same time both autonomous and geared to the needs of translation. We then face the question: what is the optimum point at which the source language and the target language should impinge on one another? Let us suppose we possess two documents: one, consisting of a descriptive analysis of each of the two languages, the other, a body of texts in the two languages, the one text a translation of the other. In the first document we find that in Language 1 there is a system A with terms n, o, p, and in Language 2 a system B with terms

q, r, s, t. The second document reveals a translation overlap between these systems such that we can make a synthesis as follows: Language 1, system A_1, terms n_1, o_1, p; Language 2, system A_2, terms n_2, o_2, q, r, where the use of the same letter indicates probability greater than a certain arbitrary figure that translation equivalence exists. Meanwhile document one has specified what are the determining features (contextual, grammatical etc.) of the two systems, and the proportional overlap between the two sets of determining features represents the minimum probability of translation equivalence. The actual probability of translation equivalence is always greater than the determining features show, because although (a) if a contextual feature X determines both n_1 and n_2, there is predictable equivalence since by definition if X is present for one text, it is present for its translation, yet (b) if n_1 is determined by a grammatical feature Y of Language 1 and n_2 by a grammatical feature Z of Language 2, there is no predictable equivalence though equivalence will arise whenever Y is found to be the translation equivalent of Z.

Since translation, although a mutual relation, is a unilateral process, what we are interested in is the choice of forms in the target language, let us say Language 2. Document one (which is presumed for this purpose to be ideal, though it must be stressed that at present there is no language which does not still require to be swept by many maids with many (preferably electronic) mops before such an ideal description is obtained) has given us the determining features of all forms in Language 2, and document two has shown us what forms of Language 2 can be predicted with what probability to be the translation equivalents of what forms of Language 1. (However ideal document two, there can never be certainty of equivalence throughout; the reason will be clear from document one, which shows that it is not the case that all languages are determined by the same features differently distributed, but that features which are determining for one language are nondetermining for another.) The final output of the translation process is thus a result of three processes, in two of which the two languages impinge upon one another. First we have translation equivalence, second, equivalence of determining features, third, operation of particular determining features in the target language. This is not necessarily a temporal order of procedure, but it may be illustrated in this way: suppose a Chinese sentence beginning ta zai nali zhu-le xie shihou giu . . . Translation equivalence might give a positive probability of Chinese non-final perfective = English simple past perfective: zhu-le = lived. (This identification is chosen for the sake of example, and is based merely on probability.) Equivalence of determining features overrules this by showing that some feature such as "past time reference relative to absolute past time" determines English past in past perfective: zhu-le = had lived. A particular determining feature of English, however, connected with the non-terminal nature of the time reference (which is irrelevant in Chinese) demands the imperfective: so we get "When he had been living there for some time. ."

Now the 'ideal' translation may be thought of as the 'contextual' one: it is that in which the form in Language 2 operates with identical effect in the identical context of situation as the form in Language 1. Theoretically, the one thing which it is not necessary to have to arrive at such a translation is the original: the first of the three processes above can be left out. But in translation in practice, one always has the original (the text in the source language), and what one does not have is the complete set of its determining features. The human translator may implicitly abstract these from the text, but this may not be wholly possible in any given instance, since the text may not contain indications of them all; and in any case the computer cannot do this until we have the complete ideal linguistic description. In mechanical translation the second of the three processes becomes the least important because it can be least well done; and the computer must concentrate on the first and the third: that is, the translation equivalence between source and target language, and the particular determining features of the latter. The less use made of comparative systematization, the more use must be made of the particular systematization of the target language. In translation as in any other linguistic composition a great deal is determined internally, by the structure of the target language; if the source language is going to yield only, or mainly, translation equivalence (as it must unless, as said above, we are to have a different description for each language in each pair in which it occurs) maximum determination must be extracted from within the target language.

For this we require a systematic description of the target language, which will be the same

whatever the source language, since it is accounting for features that are quite independent of the latter. It is quite clear what this means for the grammar: a formal grammatical analysis which covers the description of the relations between grammar and context to the extent of those contextual features which can be abstracted from the language text (not those which are dependent on situational features not themselves derivable from the text). In the example given above, we have to get both the past in past (had lived) and the imperfective (been living) from English context-grammar alone (if you try to get them through the source language text the procedure will be immensely complicated and will depend on transfer grammar, thus losing generality, since each source language will then have to have a different treatment for every target language, i.e. the Chinese of Chinese-English will be different from the Chinese of Chinese-Russian, without in any way simplifying the treatment of the target language): to get the English tense-aspect complex out of the English is relatively simple, whereas to get it out of the Chinese is absurdly complicated. There will be in other words a mechanical grammar of target English to account for the internally determined features of the language. One has only to think of source texts in Italian, Russian, Chinese and Malay to realize how much of the grammar of the English output would be left undetermined by the highest common factor of their grammatical translation equivalences.

Lexis

The problem has been discussed so far in terms of grammar, but it arises in the same way with the lexis. The first stage is likewise one of translation equivalence, the second stage is the use of the determining features of the target language. The question is: how can the lexis be systematized so as to permit the use of 'particular' (non-comparative) determining features, and especially, is it possible to operate the second stage to such an effect that the first stage can be almost restricted to a one-to-one translation equivalence (in other words, that the number of translation homonyms can be kept to a minimum, to a number that will be as small as, or smaller than, the number of historically recognized homographic (or, with a spoken input, homophonic) words in the language), which would clearly be of great advantage to the computer?

What is required is a systematic arrangement of the lexis which will group together those words among which some set of 'particular' determining features can be found to operate. Any arrangement based on orthography or phonology is obviously useless, since orthography plays no, and phonology very little, part in determining the choice of a given word at a given time. A grammatical arrangement by word classes adds nothing if, as is proposed, grammatical features are to be carried over separately as non-exponential systems, since classification is also in the main irrelevant to word determination, and where it is not, the grammar will do all that is required. (This merely amounts to saying that we cannot use grammar to determine the lexis because grammar will only determine the grammatical features of the lexis.) The form of grammatical systematization suggested above gives the clue: what is needed is a lexical arrangement with contextual reference. The lexis will be ordered in series of contextually related words, each series forming a contextually determined system, with the proviso that by context we mean (a) collocation, that is specifically word context, the statistically measured tendencies for certain words to occur in company with certain others, and (b) those non-collocational features of the context which can be abstracted from the language text.

The lexis gives us two points of advantage over the grammar, in reality two aspects of the same advantage, which arise from the fact that lexis reflects context more directly than does grammar. In the first place, one-to-one translation equivalence has a higher probability of resulting in translation in lexis than in grammar — there are whole regions of the lexis, especially in technical vocabulary, where it works with near certainty; and in the second place, where there is no 'term' (word) equivalence there is usually at least 'system' (series) equivalence. So we exploit the first advantage by giving one-to-one equivalence at the first stage, and the second advantage by the 'series' form of arrangement.

Thesaurus

The type of dictionary in which words are arranged in contextually determined series is the thesaurus. Each word is a term in one, or more than one, such series, and the translation equivalents provided by the first stage of the dictionary program function as "key-

words" leading in to the second, the thesaurus, stage. Each word will pass through the thesaurus, which will either leave it unchanged or replace it by another word in the series.

Each thesaurus entry, that is one series with its "key-word(s)", thus forms a closed system among whose terms a choice is to be made. We are already in the target language as a result of the translation equivalence of the first stage, and a pre-thesaurus output would be an interlingual form of the target language including some elements which were not words — since some key-words are in fact non-verbal symbols introduced to deal with the 'partial operator' sections of the lexis, to which we shall return later.

By the time the thesaurus stage of the dictionary program is reached we have one word in the target language (more than one word in the case of homonyms, and a symbol in the case of partial operators). We may also have a general context indicator from the source language of the type that most mechanical translation programs have envisaged, giving a clue to the generalized class of discourse in which we are operating. How much is still left to be provided from the resources of the target language itself can be gauged from a few specimens of non-technical railway terminology given below. Only four languages have been used, English, French, Italian and Chinese; and three of these are in close cultural contact; and yet there is so much overlap that we have a sort of unbroken "context-continuum" ranging (in English) from "railway station" to "coach". It is admittedly something of a tour de force, in that the words used are not the only possible ones in each case, and adequate translation would result, at least in some instances, from the use of other words. But if we consider each language in turn as a source language, each one is a possible non-translation form, and a one-to-one word equivalence would clearly not result in translation between any pair of languages, let alone among the whole four. Moreover, the sentences used were not chosen as containing words especially liable to overlap, but merely because the present writer happens to be interested in railways and in the linguistics of railway terminology.

Each sentence is given in English, because it is the language of this paper, together with a brief indication of situational or linguistic context where necessary. The underlined words, and the words in the French, Italian and Chinese lists, are contextual translations of each other: that is, words which a speaker of each language would be likely to use in an utterance having the same 'meaning' (i.e. the same place in the same sequence of linguistic and non-linguistic activity) in the same situation. They are considered as operating in a spoken text, where much of the context is situational; but in a written text, which we envisage for mechanical translation at present, the absence of "situation" is compensated by a fuller linguistic context, which is what the computer can handle. It should be stressed that, although only one word is given in each case, this is not regarded as the only possible word but merely as one which would not be felt to be out of place (this is in fact implicit in the criterion of 'the same meaning', since if it were felt to be out of place it would alter the context-sequence).

Finally, the English is British English; I do not know the American terms, but I suspect that even between British and American English there would be no one-to-one translation equivalence!

As with grammar, the systematization of the features determining the choice among terms in a lexical series requires a vast amount of statistical work, the result of which will in fact be the simplest statement of the lexical redundancy of the language. This redundancy is reflected in the fact that the terms in the commutation system operating at any given point in a context sequence are very restricted. (Two terms in a system are said to commute if one can be replaced by the other in identical context with change of meaning. If no such replacement is possible, or if replacement is not accompanied by change of meaning, they do not commute.) The restrictions can be systematized along a number of different dimensions, which will vary for different languages. The sort of dimensions that suggest themselves may be exemplified from the sentences below.

(i) Chinese huochezhan, chezhan and zhan in (2), (3) and (4) do not commute; they might commute elsewhere (e.g. huochezhan and chezhan, to a bus driver) but here they are contextually determined along a dimension which we may call 'specification', ranging from the most general term zhan to the most specific huochezhan. In mentalist terms, the speaker or writer leaves out what is rendered unnecessary by virtue of its being either

"given" in the context (linguistic or situational) or irrelevant. The computer does not know what is irrelevant — in any case irrelevance is the least translatable of linguistic phenomena — but it does know what is given, and would select zhan here if certain words are present in the context (railway terms such as huoche, and the ting (stops) of (5)), chezhan if there is some reference to a specific form of travel, and huochezhan otherwise.

(ii) English track, line, railway: the choice in (12), (14) and (16) is not a matter of specification but of classification. Like the three Chinese words, they may denote one and the same physical object; but their connotations are as it were respectively 'ential', functional

NON-TECHNICAL RAILWAY TERMINOLOGY

Situational or Linguistic Context	English	French	Italian	Chinese
1. Here's the railway station (pointing it out on a map),	railway station	gare	stazione ferroviale	huochezhan
2. How do I get to the station? (inquiry in the street).	station	gare	stazione	huochezhan
3. Station, please! (to taxi driver)	station	gare	stazione	chezhan
4. There's one at the station (on the way to the station, to companion who inquires e. g. about a post office)	station	gare	stazione	zhan
5. How many stations does it stop at? (on the Underground)	station	station	stazione	zhan
6. It's two stops further on.	stop	arrêt	fermata	zhan
7. It doesn't stop at the halts (i.e. only at the staffed stations)	halt	halte	fermata	xiauzhan
8. Travel in this coach for the country platforms.	platform	point d'arrêt	fermata	yetai
9. They' re mending the platform.	platform	quai	marcia-piede	yetai
10. He's waiting on the platform .	platform	quai	marcia-piede	zhantai
11. The train's at Platform 1.	platform	quai	binario	zhantai
12. I dropped my cigarettes on the track (while waiting at station)	track	voie	binario	guidau
13. Don't walk across the line.	line	voie	binario	tiegui
14. The trains on this line are always late.	line	ligne	linea	lu
15. There's a bridge across the line.	line	ligne	linea	tielu
16. He works on the railway.	railway	chemin de fer	ferrovia	tielu
17. I'd rather go by rail .	rail	chemin de fer	ferrovia	huoche
18. Let's go and watch the trains .	train	train	treno	huoche
19. Get on to the train! (standing on platform)	train	train	treno	che
20. There's no light in this coach .	coach	voiture	vettura	che

and institutional. A purely locational context could give 'track', a proper name 'railway'; 'line' overlaps with both (cf. (13) and (15)) and might be limited to functional contexts such as 'main line'.

The word as a term in a thesaurus series is grammatically neutral: it is neutral, that is, as to all grammatical systems, both categories of the word (e.g. number) and word class itself. Since we cannot carry over the classes and other categories of the source language as one-to-one equivalences (e.g. Chinese verb ≠ English verb, Chinese plural ≠ English plural, even if both languages are described with categories named 'verb' and 'plural'), these are dealt with in the grammatical part of the program and only after having reached the target language do they re-enter the range of features determining word choice. The attempt to handle such categories lexically leads to impossible complexity, since every word category in each source language would have to be directly reflected in the thesaurus.

All mechanical translation programs have carried over some word categories non-lexically, word-inflections obviously lending themselves to such treatment. If in the thesaurus program the word is to be shorn of all grammatical features, including word class, the whole of the grammar must be handled autonomously, and the method proposed for this is the lattice program originated and developed by Margaret Masterman and A.F. Parker-Rhodes. The lattice program, which is a mathematical generalization of a comparative grammar (i.e. a non-linguistic abstraction from the description of a finite number of languages) avoids the necessity of the comparative (source-target) identification of word (and other grammatical) categories. The word class of the target language is determined by the L(attice) P(osition) I(ndicator), derived from the grammar of the source language; class is thus not a function of the word as a term in the thesaurus series, nor does the choice of word class depend on comparative word class equivalences.

The autonomy thus acquired by the lexis of the target language allows the thesaurus stage of the dictionary to be the same for one target language whatever the source language, and at the same time permits the maximum use of the redundancy within the target language by allowing different treatment for different sections of the lexis. This would be impossible if word classes were based on translation equivalence, since the thesaurus series could not form closed systems within which determination can operate. If for example one identified particularly (i.e. non-comparatively) a word class 'conjunction' in the target language, the redundancy of the conjunction system can only be fully exploited if it is determined (as it is by the LPI) that the choice word must be a term in this system. If we attempted to carry over to Chinese word classes from, say, English, where we could not identify any grouping (let alone class) of words which would have valid translation equivalence with Chinese 'conjunction', we should forfeit the redundancy of the Chinese system since the words among which we should have to choose could not be ordered as terms in any lexical series.

The thesaurus admits any suitable grouping of words among which determination can be shown to operate; the grouping may be purely lexical or partly grammatical (i.e. operating in the grammatical system of the target language). It might be that a word class as such, because of the redundancy within it, was amenable to such monosystemic treatment. This is clearly not the case with the 'non-operator" (purely lexical) sections of the lexis, such as verbs and nouns in English, but may work with some partial operators. (Pure operators, i.e. words not entering into lexical systems, which are few in any language (since their work is usually done by elements less than words) — Chinese de is an example — will not be handled by the thesaurus, but by the lattice program.) The nouns in the above sentences enter into lexical series, but no determination system can be based on their membership in the word class of 'noun'; prepositions, on the other hand, which are few in number — and of which, like all partial operators, we cannot invent new ones — can in the first instance be treated as a single lexical grouping.

It is simply because partial operators (which in English would include — in traditional 'parts of speech' terms — some adjectives (e.g. demonstratives and interrogatives), some adverbs (those that qualify adjectives), verbal operators, pronouns, conjunctions and prepositions) are in the first instance grammatically restricted that they have a higher degree of overall redundancy than non-operators. Knowing that a noun must occur at a certain point merely gives us a choice among several thousand words, whereas the occurrence of a verbal operator is itself highly restrictive.

An idea of how the thesaurus principle might be applied in a particular instance may be given with respect to prepositions in English. In dealing with the English prepositions we can begin by considering the whole class as a lexical series. We can then distinguish between the 'determined' and the 'commutable'. Most prepositions are determined in some occurrences and commutable in others. The 'determined' prepositions are simply those which cannot commute, and they are of two types: the pre-determined — those determined by what precedes (e.g. 'on' in "the result depends on the temperature at . . ", which cannot be replaced, or 'to' in " .. in marked contrast to the development of . .", which could be replaced by 'with' but without change of meaning), and the post-determined — those determined by what follows (e.g. 'on' in "on the other hand", or 'to' in "to a large extent"). In the system of each type we may recognize one neutral term, pre-determined 'of' and post-determined 'to'.

Determined prepositions will be dealt with not as separate words but as grammatical forms of the word by which they are determined. The combination of pre-determining word plus preposition will constitute a separate entry, a transitized form of the determining non-operator (verb, noun or adjective, including adverb formed from adjective), of which the occurrence is determined by the LPI. The features determining the occurrence of these forms are grammatical features of the determining word; they are connected in varying ways with the presence or absence of a following noun (group): 'depends / depends on A', 'a contrast / a contrast with A', 'liable to A'; but 'wake up / wake A (up)'. Which form of the word (with or without preposition) corresponds to which lattice position will be indicated if necessary in the same way as other word class information; in the absence of such indication the transitized form of words which have one is used before a noun. If a verb is not assigned a marked transitized form, it is assumed not to have one, and will be left unaltered in a lattice position that would require a transitized form if there was one; but if a noun or adjective without transitized form occurs in the corresponding lattice position the neutral term 'of' is to be supplied. Thus 'depend', 'contrast (noun)' have the transitized forms 'depend on', 'contrast to'; 'display', 'production', 'hopeful' have no transitized forms, and will thus give 'display of (power)', 'production of (machinery)', 'hopeful of (success)'.

Post-determined prepositions are always treated as part of a larger group which is entered as a whole. These are forms like 'at least', 'on the whole', 'to a large extent', and are single words for thesaurus purposes. The exception is the neutral term 'to' before a verb (the 'infinitive' form). This is treated as a grammatical form of the following word (the verb) and will be used only when required by the LPI, e.g. in a two-verb or adjective-verb complex where the first element has no predetermined (or other) preposition: 'desires to go' but 'insists on going' — all other prepositions require the -ing form of verbs —, 'useless to go' but 'useless for (commutable) experiment'.

Determined prepositions in the English version of the Italian pilot paragraph are:

Pre-determined:	of 1 - 6
Post-determined:	at least; on the other hand; in fact; for some time past; above all; to mechanize.

Commutable prepositions operate in closed commutation systems of varying extent (e.g. 'plants with/without axillary buds' (two terms only), 'walked across/round/past/through/towards etc. the field'), and each one may enter into a number of different systems. Those which are lexical variants of a preceding verb are treated as separate lexical items, like the pre-determined prepositions (e.g. 'stand up', 'stand down', and favorites like 'put up with'). The remainder must be translated, and among these also use is made of contextual determination.

The overlap in this class (i.e. among words in source languages which can be translated into words of this class in English) is of course considerable, as one example will show:

Sentences:	English	Italian	Cantonese
He went to London	to	a	
He lives in London	in	a	hai
He came from London	from		hai

We can however set up systems limited by the context in such a way that the terms in different systems do not commute with one another. For example, concrete and abstract: to / in / from commute with each other but not with in spite of / for / without. Within the concrete we have motion and rest: to / from commute with each other but not with at / on / under; and time and place: before / after / until

commute with each other (in some contexts before / until do not commute but are grammatically determined) but not with under / at.

Commutable prepositions of this type will go through the usual thesaurus program in which they form series on their own (whereas determined prepositions and the 'lexical variant' type of commutable prepositions do not); the context will specify in which system we are operating. If the source language has words to which English prepositions are given as translation equivalents, these will as usual be one-to-one (with limited homonymy where necessary: Cantonese hai would have to give 'be at (English verb or preposition according to LPI); from (preposition only)', since on grounds of probability the motion context equivalent of 'at' will be motion towards, not away from). Each key-word will in the usual way lead into a series the choice within which will be determined by the context category.

Commutable prepositions in the Italian pilot paragraph are:

Lexical variants:	none	
Free commutables:	with	(It. a, abstract 'with (/without)'
	for	1 - 4 (It. per, abstract)
	in	(It. in, abstract)

This paragraph is typical in that the freely commutable prepositions are a minority of the total prepositions in the English output.

Thus the thesaurus method,which uses the contextual determination within a language, is applicable to partial operators through the handling of redundancy at the level at which it occurs: where the use of a preposition depends on grammatical or lexical features (considering English forms like 'put up with' to be lexical, not contextual, variants) it will be handled accordingly, and not as a term in a lexical preposition series. The method is far from having been worked out in full; the principle on which it rests, that of "make the language do the work", can only be fully applied after the linguists have done the work on the language.