[From: *Zeitschrift für Phonetik* 15(1) 1963]

M. A. K. HALLIDAY, EDINBURGH

Linguistics and Machine Translation

Linguists often like to speak of linguistics as a science. It is true that each one of us may have his own view of exactly what that implies; but it must mean at least that linguistics is to be scrutinized according to some general criteria by which scientific theory and scientific methods are assessed. We should like linguistics to satisfy these criteria, not because the term "science" is a kind of status symbol for an academic subject but because we want to be able to account for linguistic events in a way that is, and can be shown to be, both valid and useful. In the "pure" side of the subject, that known frequently as "general linguistics", what is being sought is a valid and powerful theory, and rigorous methods; while for applied linguistics we need to be able to say things about language that are useful in any field of activity where the study of language is central. It is fair to say, I think, that linguistics has gone some way towards both these goals.

One important area for the application of linguistics is machine translation (for which I will use the affectionate abbreviation MT). Rather surprisingly, it has not always been obvious that MT is, among other things, applied linguistics. In part, this has been because linguistics, or at least that side of linguistics that is relevant to MT, was a fairly new subject without other known applications; in part it is due to the history of MT itself. It is convenient to date MT from the famous 1949 memorandum by WARREN WEAVER; but its origins include such sources as wartime work on code-breaking, the theoretical study of communication and the development of information theory, and the design and construction of electronic digital computers. It was probably these origins that gave to the early work on MT its particular slant: most of the effort was directed towards problems in electronic engineering and in the coding and transmission of information. This in turn determined the approach of many of those working in the field.

The solution of the relevant engineering and mathematical problems is of course a prerequisite of success in any MT project. But there is another side to the subject, as recognized already in the discussion at the VIIth International Congress of Linguists in London in 1952. This is the application of theory from linguistic science. MT is a problem in applied linguistics, specifically a problem requiring the application of those parts of General Linguistic theory which deal with the systematic description and comparison of languages: descriptive linguistics (known often in America as "synchronic linguistics" and by some people on both sides of the Atlantic as "structural linguistics"), and its recent derivative, comparative descriptive linguistics.

This body of theory is neither "an approach" nor "just common sense", any more than it was common sense which sent Major Gagarin into space, but a specialized scientific theory which, like any other of its kind, has to be learnt. From it are derived the methods for the description and comparison of languages, methods demanding rigorous and consistent application. Since the theory has been worked out and applied by linguists working on a large number of different languages in different parts of the world there are, as in a lot of subjects, divergent "schools"; but the area of agreement among them is far wider and more fundamental than are the divergencies.

The name "linguistics" reflects the independent and autonomous status of the subject as the science of language, that is to say, its categories and methods are not drawn from philosophy or logic or psychology or anthropology or mathematics or communication theory, but from language itself: from, that is, a theory founded on hypotheses that were set up to explain the way language works. Language, for this purpose, is regarded as a unique, patterned form of human social activity, operating in the context of other social activities; and the job of descriptive linguistics is to state the patterns of a language and show how it works. This inevitably gives the linguist a standpoint different from that of his colleagues in any of the other disciplines which study language as part of the material of their own research: his object is to throw light on language, not to use language to throw light on something else. This particularity of the linguistic study of language is especially marked in the study of meaning: this is an essential and inseparable part of linguistic description and is not the same thing as, for example, the psychological or philosophical study of meaning.

Work on MT has brought out very clearly the contrast between the linguistic view of language and some of the views arrived at by looking at language through the lenses of other disciplines. Two opinions current in MT writings on language are that language is a code and that the code is fundamentally binary. Both these views are, from the standpoint of a communication engineer, tenable and useful. From the linguistic standpoint, however, these views are both questionable and unhelpful; and they have hampered MT work because they misrepresent the functioning of language both in its internal relations and in its relations to non-language. A code implies encoding and decoding processes, and a systematic relationship between code and message, which must be a relation between two distinct entities since the message exists independently of the code and does not determine the type of the code. Language however does not yield this dichotomy: in language, code and message are one, and there is no relation between two entities one of which can be chosen independently of the other. One may force an analogy, but I do not think it has been shown to be a fruitful one, between the categories of code and message on the one side and those of form and content (or concept) on the other; but concepts are not accessible to study except through linguistic forms and are not present as independent parts of the material, the '"text", that the linguist — or any other student of language — has to work with. The one linguistic process which could be regarded as a special instance of the coding process is that of translation, where a language Ly, the target language, can be said to stand in the relation of a code to a language Lx, the source language; but this is an entirely different point and is clearly not what is meant by those who say that language is a code.

Nor, to take the second point, do the patterns of language operate exclusively in binary oppositions. Grammatical systems, like all closed systems, can of course be stated in binary terms, and this is a convenience for statistical and electronic linguistics. But it is not a general truth about language, and to treat it as such will complicate any piece of work in which language data are being processed. If it is to be capable of rigorous definition and application to natural languages, a linguistic system, which is a pattern of a particular type with certain specific characteristics, must be allowed to have any number of terms. Language is like that. The attempt to force all language systems into binary mould has required recourse to the extreme of saying that the choice of any linguistic item is in binary opposition to the nonchoice of the same item; which is of some significance as a logical proposition but to linguistics is an irrelevant platitude. There happens to be one type of binary opposition that is fundamental to language, that of marked and (formally or contextually) unmarked terms in a grammatical system, or indeed in an open lexical set: an example from grammar would be English

	formally	**contextually**
"student"	unmarked	marked (= singular)
"students"	marked (+ s)	marked (= plural)

which contrasts with Chinese

"xuesheng"	unmarked	unmarked (= singular or plural)
"xueshengmen"	marked (+ men)	marked (= plural).

But not all systems have an unmarked term, and those that do often have more than one marked term; so that this feature, important as it is, is not by itself adequate as a primary dimension for the definition and description of language systems.

To describe a language meaningfully it is necessary to replace the vague generalizations and random observations transported ad hoc from other disciplines with a consistent and comprehensive theory of language; the methods used in description are derived from, and answerable to, this theory. This is no new idea: descriptive linguistics in ancient India was already a specialized study in its own right, and attained a sophistication of theory and rigour of application which were not surpassed in Europe until the present century, and which some descriptions used by learners of languages still fall far short of. It is impossible to describe language without a theory, since some theory, however inadequate, is implicit in all descriptions; but it is quite possible to make a description that is in practice unsystematic, with categories neither clearly interrelated nor consistently assigned. Such a description, however, is difficult to programme on a computer, since the computer has no native language (and has had no classroom Latin) to which to relate an agglomeration of ad hoc categories. If the computer is to be expected to translate, then it is all the more essential that it should be provided with descriptions that are both accurate and theoretically valid, since it has to digest not only the descriptions of two languages, but also the rules for the systematic relating of these two descriptions one to the other. It is expected in fact to perform a complex operation of comparative descriptive linguistics, of which translation can be regarded as a special instance.

There are a number of things that can be said about translation from the linguistic point of view. For this it will be necessary to introduce some categories from linguistic theory, the technical terms for which are usually disguised as ordinary words, like "level" and "unit" and "form" and "rank" — linguists, especially European linguists, tend to prefer redefining old terms to coining new ones. "Levels" are the different types of pattern found in language: the two major types are "form", the meaningful organization of language, and "substance", the raw material of language; but the total range of levels, with subdivisions, is more complex. We need two levels of form, "grammar" and "lexis", two of substance, "phonic" and "graphic", two levels for patterns of the arrangement of substance in form, "phonology" and "graphology" (the two latter pairs, unlike grammar and lexis, being in "either/or" relationship, since language text is either spoken or written), and one level for patterns of the reflection by language of things that are not language, namely "context". Form and context are the principal levels for the statement of meaning; form is the internal aspect of linguistic patterning, the relation of the parts to each other, context the external aspect, the relation of the formal items to other features of the situations on which language operates. Within form, grammar is the operation of items in closed system contrast, characterized by very complex relations with a small number of terms; lexis the operation of items in open set contrast, very simple relations with large numbers of terms. Much more abstraction is therefore possible in the statement of grammar than in the statement of lexis; in grammar we can recognize not only grammatical "items" but also abstract categories such as "units", the stretches of differing extent which carry patterns, "structures", the patterns of arrangement, and "classes", the groupings of items that behave alike. In lexis, on the other hand, only the items themselves enter into patterned relations.

Translation, as a process, is unidirectional; but a translation, which is the end-product of such a process, is in mutual relation with the original: either of the two texts could replace the other as language activity playing a given part in a given situation. Taken together the two texts constitute a type of comparative description of the two languages, in which the two languages impinge on each other at a number of different levels. We can leave aside phonology and graphology, since these relate the form of a language to its spoken or written substance and are only accidentally relevant to translation within certain pairs of languages. For translation the important levels are those of form, namely grammar and lexis, and that of context. At these levels we have as it were two types of evidence, or two bodies of source material, for the comparison — that is, for the bringing into relation with one another — of the two languages.

On the one hand, there are actual translations; texts in the two languages, the one being translated from the other. These display probabilities of equivalence between items occurring in them. Such probabilities may be stated as simple, unconditioned probabilities, or as conditioned either in sequence or in combination. So one can ask, for example, "what is the most probable tense of the English verbal group in translation of the Russian past tense?"; or what is the most probable tense of the English verbal group in translation of the Russian past tense if the latter is (a) in a clause preceded by a clause in which the verbal group also had past sense, or (b) combined with perfective aspect?" And of course one can state such probabilities in gradation, from most to least probable, and in quantified form; and one can take into account sequences and combinations of three, four and more features.

On the other hand, there is the material furnished by a comparative analysis of the two languages, in which their formal and contextual features are described by means of a single set of categories. Since descriptive categories are not universals — just because we find something we want to call a "verb" in one language this does not necessarily mean we shall find such a thing in another language — categories used for the comparison of two languages are drawn from the separate description of each. When they are conflated into a single set, degrees of likeness can be stated and measured. By stating the formal and contextual equivalents of grammatical structures and items, and of lexical items, with one set of categories, the comparison brings out the likeness and unlikeness between the two languages. This takes account also of partial equivalence, where correspondence is not one to one but systematic statement is still possible.

For lexis the statement of equivalence between two languages is traditionally made by bilingual dictionary. The aim of the ideal dictionary could be said to be to state under what circumstances a word in Ly is the contextual equivalent of a word in Lx. The bilingual dictionary faces, in the main, two difficulties. One is the nature of "equivalence": the contextual equivalent is not necessarily the translation equivalent, since the latter is determined not only by contextual meaning but also by formal meaning, that is by the tendency to occur in collocation with other words. This difficulty can be met by extending the concept of a citation to cover not just *a word* in a given collocation but *a collocation* of words: the item for the statement of equivalence would not be "to climb (mountains &c.)", since the contextual equivalent of "climb" here might not in fact collocate with all the words that are the contextual equivalents of "mountain" (still less with those of the "&c."), but would rather be "to climb a mountain".

The other difficulty is that the category of "word" is not a universal constant. The term "word" is in fact highly misleading. In the first place, even within one language it is used to cover, and thus to conflate, two and sometimes three quite different categories, whose members only partially coincide. The name "word" is given to one of the units in grammar; this unit is then taken as the unit for dictionary entry. But the dictionary exists to state lexical, not grammatical, meaning; what the dictionary should be describing is the "lexical item", which is not always coextensive with the grammatical word. In those languages, such as English, which also have an orthographically defined "word" — defined as "that which occurs between two spaces in the script" — the "word" is now a unit on *three* levels and the probability that all three will exactly coincide is even lower. In the second place, even when the dictionary is freed from the tyranny of the grammatical word and allowed to handle the unit that is appropriate to it, the "lexical item", this latter is not a constant across languages. Languages vary very much in their organization of lexical meaning — that is, in how much and what they put into the items that operate in open sets and in collocation relations. This makes it all the more important in comparative description that the lexical item should be properly identified and its meaning stated according to how it works in the language. Since in a dictionary the order of items is linguistically almost random, for comparative purposes the thesaurus, which groups lexical items in their sets, may be a more useful method of lexical description.

In grammar we have as yet no complete comparative descriptions of any pair of languages, since these cannot be produced without prior statistical descriptions of each language, which require facilities that linguists are only just beginning to secure. Such descriptions are now both necessary and possible: necessary to the more effective *application* of linguistics, in MT and elsewhere, but made possible by advances in linguistic *theory.* The distinction is important: the move from qualitative to quantitative description is of no value whatever unless description is anchored in sound and scientific theory. This is why so much work that has gone into the counting of orthographic words in isolation, or into the laborious compilation of tables of occurrences of ill-defined, shifting, supposedly universal and often entirely non-linguistic categories, is sadly unrewarded. No doubt the uselessness of such work has influenced those MT workers who have denied the value of statistical descriptions of language. But once it is realized that linguistic theory can ensure that statements about language are meaningful, and thus turn MT from the hit-and-miss subject it has tended to be into a scientific discipline, such descriptions fall into place. It is not merely that statistical descriptions are needed for future progress in linguistic theory and method; they are of direct value in application. This is true even in the application of linguistics to language teaching: for example, all textbook accounts of the English so-called "phrasal verbs" are unsatisfactory, their operation being such that only quantitative analysis will yield an adequate classification. How much the more is it true of MT: translation is essentially a "more likely less likely" relation, and if a computer is to translate adequately it cannot operate on "yes no" evidence alone.

Grammatical equivalence between two languages can be displayed most adequately, therefore, by means of quantitative studies of the grammar of each. Such equivalence must furthermore be related to the "rank"' scale: the scale of grammatical units, of which the "word" is one. These units are the stretches into which language text is cut when grammatical statements are being made about it. Again they are not universals: they must be recognized afresh for each language. When we compare two languages we cannot link the languages as a whole; we select for comparison items from within them — and not only items, of course, but abstract categories (classes, structures and so on) of which the items are "exponents". These items, and the categories set up in abstraction from them, must be related to the grammatical units of which they are members.

So for comparative purposes we need first to relate the *units* of the two languages to each other on the basis of probability of translation equivalence. If we can say, for example, that a "sentence" in Lx can usually be translated by a sentence in Ly — this being implied when we call the two units, one in Lx and the other in Ly, by the same name — then we can try to make a systematic statement to account for the occasions when this does not work. Suppose for illustration that we can describe Lx and Ly, separately, each with a system of five grammatical units, which we will call, in descending taxonomic order, "sentence, clause, group, word, morpheme". Then a clause, for example, in Lx will normally, but not always, be translated by a clause in Ly. Grammatical equivalence, in a comparative description, can be sought and stated at the rank of any one of these units: each must be taken by itself, since each has its own structures, classes and systems through which to display the formal similarities and differences between the two languages, and its own set of items as possible translation equivalents. If Lx, for example, makes a class distinction at the rank of the group between verbal group and nominal group, does Ly make the same distinction? And if so are the items which are exponents of the verbal group in Lx always translated by items which are exponents of the verbal group in Ly?

Lexical equivalence is not to be sought exclusively at any rank in the grammar. While the reason why, in the grammar of any language we call one of the units "word" is that that unit, more than any other, yields lexical items, what defines the lexical item is not the fact that it is grammatically a word — it may not be — but the fact that it cannot be fully described by the grammatical categories. The latter account for the "closed systems" of language, and the items entering into them, this being the essential characteristic of grammar. One important consequence of the difference between grammar and lexis is that information theory, which is entirely appropriate to the statement of grammatical meaning — since information is a property of closed systems — is at present no use for the study of lexis: there is no way of quantifying the information carried by an open set.

As an illustration of the translation process, below are given two sentences, each in both a Chinese and a Russian version. These sentences are shown segmented into their successive lower units: clauses, groups, words and morphemes, with each boundary implying all those below it (a group boundary must also be a word and morpheme boundary, and so on). The segmentation has been made to display maximum likeness, as in a comparative description of the two languages; it is also oversimplified in two ways: first, lineally discrete segments have been recognized throughout, though in fact grammatical units overlay one another (e.g. the English word "ran" consists of two morphemes, but these do not occur as one following the other), and second, "rank-shifting" has been avoided where possible. "Rank-shift" is the operation of one unit in the structure of a unit of lower rank: (e.g. a clause by definition operates in sentence structure, but in "the man who came to dinner", "who came to dinner" is a rank-shifted clause operating inside a nominal group).

Each sentence is then "translated" at each rank into English: first each morpheme is taken separately, then each word, and so on. In each case the English equivalent *item* is one which might turn out — at a guess: the counting has not been done — to be the most frequent translation equivalent *at that rank:* the one which would be the first choice for entry in a bilingual "dictionary" of morphemes, words, groups &c. Similarly the grammatical *pattern* chosen is that which might be the most frequent translation equivalent at the rank concerned. (The concept "most frequent translation equivalent" for a *grammatical item* in isolation, such as English "the" or "-ing", is however inapplicable; such items are here symbolized "X" until their incorporation into higher units.) If we start from the morpheme, we can follow the translation step by step up the rank scale, each equivalent being adjusted as it finds itself co-occurring with certain other items, in a certain grammatical relation, in the unit next above. So for example Chinese *tie,* as a morpheme, would most frequently be translated "iron"; when it is taken as part of the word into which it enters, this translation is the one most likely to appear (as when it is a word on its own, or in the words *tieqi* "ironware" or *shengtie* "cast iron"); elsewhere other equivalents must be chosen *(gangtie* "steel", *tielu* "railway").

Each step can be regarded as a process in which the equivalent is retained unless positive contrary indications are found in the next unit.

It appears clearly that, while equivalence can be stated, in terms of probabilities, for all ranks, translation in the accepted sense does not occur below the rank of the clause, and a good translation needs to be based on the sentence as its unit. So-called "literal" translation is, roughly, translation at group rank, or at a mixture of group and word.

Theoretically, it would be possible for MT to proceed by means of a sentence dictionary, all possible sentences of Lx being entered, together with their translation equivalents in Ly. Since, in the scientific and other registers that we are interested in translating, a sentence is hardly ever repeated in its lexicogrammatical entirety, this is a practical impossibility. By the separation of grammar from lexis it becomes, with present day computer storage, at least conceivable. Sentences containing identical sequences of lexical items (but differing in grammar) might recur; and sentences with identical sequences in grammar (but differing lexically) certainly do: regularly in the sense of "having the same primary grammatical *structure",* at least down to the rank of the word; even the same sequences of grammatical *items* probably turn up now and again.

The illustration below shows, for one of the English sentences, the grammatical and lexical material separated out. I (a) is a linear statement of the grammatical structures at all ranks; I (b) shows the grammatical items which are the exponents of each of the elements of structure. II gives the sequence of lexical items. From the point of view of linguistic theory, such separation is quite justified: indeed grammatical description and lexical description must proceed independently at first, since different relations (and therefore different theories) are involved — though description is not complete until the two levels have been related. The weakness from the point of view of MT, however, would be that in translation there must be constant cross-reference between the grammar and the lexis, since in all languages some grammatical items can only be identified by reference to lexical ones, and vice versa. For example, only the grammar of the English sentence shows which of a number of lexical items "part" is; conversely only the lexical identification of "part" allows us to say whether it is singular noun or plural verb.

In general however the unit selected as the basis for MT has been way down the rank scale, far below the sentence: usually the word or the morpheme, or a combination of both (especially where the source language is Russian). The use of the word or morpheme yields, by comparison with the sentence or clause, inventories of manageable size. At the same time it involves complex programmes for the selection among possible equivalents, usually based on items in the immediate environment (though where grammar is involved *structures* are far more powerful, because more general), plus routines for reordering the components of the units above the translation unit. So for example the morpheme word *chang* must be identified both lexically, as translatable by "long" — e.g. by collocation with *gong-li* "kilometre"; and grammatically, as a finite intransitive verb "is/are long" — which can be shown by its association with the item *gong* "altogether", a member of a small class of words which can only precede a finite verb, but is more usefully shown (since this will cover a much greater number of instances) by the identification of the clause structure.

In fact there is no reason why any one unit should be taken as the sole basic unit for translation. In describing a language we give no special priority either to the morpheme or to the sentence: all units are equally "basic". In translation too we can handle each of the units, and "shunt" from one to another, taking advantage of the highly specific way in which the units can be shown, in the theory of grammar, to be interrelated.

If we analyze the translation process with "translation equivalence" regarded as linked to the grammatical rank scale, we can distinguish three stages in it. These are not of course discrete steps taken one after the other, but rather abstractions useful to the understanding of the translation process and of "a translation" as its end-product. First, there is the selection of the "most probable translation equivalent" for each item at each rank, based on simple frequency. Second, there is the conditioning effect of the surrounding text in the source language on these probabilities: here grammatical and lexical features of the unit next above are taken into account and may (or may not) lead to the choice of an item other than the one with highest overall probability. Third, there is the internal structure of the target language, which may (or may not) lead to the choice of yet another item as a result of grammatical and lexical relations particular to that language: these can be viewed as brought into operation similarly by step-by-step progression up the rank scale. Stage three is purely descriptive; the source language no longer plays any part here. The weighting of these (descriptive) factors from the structure of the target language against the (comparative) factors drawn from the source language is one of the major theoretical problems, analogous to the weighting of input against internally conditioned probabilities in automatic speech analysis.

As an example, consider the translation of the item *duo* in the Chinese version of the first sentence below. As a morpheme it is most likely to require, in written English, the rendering "many", though there are a number of other possible equivalents. When it turns out, by reference to the unit next above, that *duo* is here a complete word, not part of a word, it becomes more likely that it is a verb, to be rendered "are many". This version is clearly unlikely to survive for long, and in many examples would be replaced at clause rank by "there are many", on internal grounds: English would transform "the problems are many" into "there are many problems". In this example, however, when we go one further up the rank scale, the place of *duo* in group structure shows that it stands to the numeral in a relationship rendered in English by the addition of "than": "many than 23,000". The rules of English require that in the structure of which this is an example the comparative form (which has no *item* equivalent in Chinese) should be selected: "more than". A more sophisticated programme might alter this at clause rank to "over" but this could not be generalized to all such occurrences of *duo:* "over three-o'clock" is not acceptable. What is as it were left over to stage three will in any case depend on the comparison of the target language with the source language: if one was translating *trois jours* from (written) French, the English plural form "days" could be arrived at by translation from *jours,* whereas in the translation of the Chinese equivalent *san tian* the use of the plural form "days" in concord with the numeral form "three" would appear as an internal feature of English.

The human translator performs all stages of the operation at all ranks in a single process. A computer programme, if it is to achieve a reasonable degree of generality even in one pair of languages in one direction, may have to handle them separately. Whether it does so or not, in either case it will need to shunt up and down among the different units of the rank scale as the translation advances towards the final choice of equivalent. Units and their grammatical structures have to be identified; and this means that a lot of prior linguistic research needs to be done. For a long time it was not recognized by many workers in MT that description must precede translation: that the main function of computers for the first few years was to produce detailed descriptions of the languages between which it was proposed to translate. This is now well known, and we have centres of machine research where descriptive work is being done which will make large-scale MT a possibility in practice, as it always has been in theory.

The studies required are on these lines. First, particular descriptions, including quantitative statements, of each of the languages concerned: both (a) formal statistics, especially grammatical — e.g. occurrences of classes, and of sequences of classes, of each unit, but also lexical — occurrences of lexical items *in collocation* (most linguistic statistics to date has been lexical and has ignored sequences) and (b) statements linking the formal occurrences to the contextual meanings of the forms concerned (contextual meaning itself cannot be stated in quantitative terms, but — since it is dependent on form — it requires for MT to be based on statistical analysis in grammar and lexis). Second, comparative descriptions, of pairs or groups of the languages concerned: either (a) direct comparison, suitable for a programme concerned with only one pair of languages such as Russian-English, or (b) indirect comparison via a machine interlingua, for a general programme to be adapted, with insertion of appropriate dictionaries, to an increasing number of different languages. The machine interlingua would be not a natural language nor an artificial language but a mathematical construct serving as a transit code between any pair of natural languages. It would of course have no output, and would reduce the total number of programmes very considerably — for example from 90 to 20 if it was desired to translate each way between each pair from among ten different languages. The interlingua approach, though not generally favoured in the past, has much to recommend it for long-term application, giving as it does more scope for exploiting the power of linguistic theory. It is also likely to be of great benefit to that theory, since it could yield — not universal descriptive categories, which would be so weak as to be useless, but — a general frame of reference for the comparative categories which have to be set up when two languages are formally compared. It is not only MT which needs comparative descriptions; they are essential to other applications of linguistics, and nowhere more than in foreign language teaching. So far, however, only MT has had the necessary facilities. It would be a pity if specialists in this important subject failed to make maximum use of the achievements of all its contributory disciplines.

Appendix

A. "Rank by rank" English translation of Chinese and Russian sentences.
Conventions: —

///	sentence boundary
//	clause boundary
/	group boundary
(space)	word boundary
—	morpheme boundary
[[ ]]	boundary of rank-shifted clause
[ ]	boundary of rank-shifted group
1	morpheme equivalents
2	word equivalents
3	group equivalents
4	clause equivalents
5	sentence equivalents
X	grammatical item
(PN)	proper name

	/[Zhong-guo	di]	tie-lu	gong	chang	/
1	(PN) country	X	iron way	altogether	long	
2	China	X	railway	altogether	is are + long	
3	of China		railway	is are altogether long		

4 the railways of China are altogether more than 23,000 kilometres in length

5 The railways of China are altogether more than 23,000 kilometres in length, of which the greater part is in the Northeast Provinces

2	wan	3	qian	duo	gong-li	//
2	ten + thousand	3	thousand	many	metric mile	
20 + thousand		3 + thousand		are + many	kilometre	

more than 23,000 kilometres

qi	da	bu-fen	shi /	zai	dong-bei	//
thereof	great	part share	X	at	east north	
thereof	is + great	part	X	is + at are	northeast	
the greater part thereof			X	is are	in the northeast	
the greater part thereof		is are	in the northeast			

	Obšč-aja	dlin-a	železn-ych dorog	Kita-ja
1	general X	long X	iron X way	(PN) X
2	general	length	of + iron of + ways	of + China
3	the overall length		of railways	of China
4	the overall length of the railways of China is over 23,000 kilometres			
5	The overall length of the railways of China is over 23,000 kilometres, of which the greater part is in the Northeast Provinces			

ravn-a	23	[	s	ličn-im	]	tysjač-am
equal X	23		with	extra X		thousand X
is + equal	23		with	with + extra		to + thousand
is equal	to 23		over			thousand

kilometr-ov	//	bol'š-aja	ich	čast'	/	na-chod-it-sja	/
kilometre X		great X	their	part X		on X X go	
of + kilometres		great	their	part		is + found	
kilometres		the greater part of them				is found	
		the greater part of them is in the Northeast Provinces					

v	provinci-jach	/	Sever-o-vostok-a	///
in	province X		north X east X	
in	provinces		of + northeast	
in the provinces			of the Northeast	

///	Wo-men	suo kan-jian di	//	bi-jiao	/
1	I X	X look see X		compare compare	
2	we	X see X		compare	
3	we	what see which		compare with	
4	what we see			compared with what we	
5	What we saw was even more interesting than			what we had heard before	

wo-men	yi-qian	suo	ting-jian	di
I X	X before	X	listen see	X
we	before	X what	hear	X
we	before	which	hear	

heard before

geng	you	yi-si	
X	have	mean	think
still + more	have	significance	
have even more		significance	
is			
are even more interesting			

	vs-jo		to	//	č-to		my	u-vid-el-i			
1	all	X	that		what	X	we	by	see	X	X
2	all		that		what		we	saw			
3	all that				what		we saw				
4	all that				what we saw						
5	All that we saw was far more interesting than what we had heard before										

gorazd-o		interesn-eje	/	to-go	
good + at	X	interesting	X	that	X
far		more + interesting		of + that	
is					
are far more interesting				of that	
is					
are far more interesting than that					

[[č-to	/	my	slyš-al-i		/	ran-se	]]///
what	X	we	hear	X	X	early X	
what		we	heard			earlier	
what		we heard				earlier	
what we heard before							

B. Grammatical and lexical features of an English sentence.

/// The railway-s [of [China]] are altogether /
more [than [23,000 kilo-metre-s [in [length]]]] //
[of [which]] the great-er part / is / in [the North-east Province-s] ///

I (a): Linear statement of sentence, clause and group structure

//α/S dhq (/pn (/h)) /P h /A h /C e = hq (/ pn
(/ohq (/pn (/h)))) *//β* / B = S q (/pb = n (/d = h)) deh
P h /A = C pn (/dsh) ///

I(b): Sequence of grammatical items

the ( )s of ( ) are ( ) more + than (numeral) ( )s in ( )
of which the ( )er ( ) is in the ( ) ( ) ( )s

II: Sequence of lexical items

railway China altogether 23,000 kilometre length great
part northeast province