[*Eighth Annual Round Table Meeting on Linguistics and Language Studies*, Georgetown University, 1957]

Meaning In Relation To MT

MARTIN JOOS
University of Wisconsin

It is not my purpose here to compete with or to anticipate the contributions which other panel members will make with regard to such things as vocabulary size and storage, or even the problem of multiple meanings in present-day programming-research, for the bearing of my paper upon such things will be thoroughly abstract. It must be perfectly clear why this paper was placed in the lead-off position on the present conference: the management wanted to start with a clean slate. There just wasn't anybody else with less experience in MT, for mine is identically zero. I have had M experience — I have even built primitive computers — and I have done all kinds of T except only the M kind of T. Can I claim that I was doing MT when, in a certain national emergency, I spent dozens of hours translating Finnish into English with dictionary and grammar alone — that is, without knowing any Finnish? On second thought, I suppose that would be at least MT plus editing, for I never considered a sentence translated until I had made good sense out of it.

MT plus post-editing? No, not quite that either, for the editing didn't come after the MT. The two things were mixed, either done alternately (with several alternations within each sentence) or else simultaneously. I speak of 'simultaneous' MT and editing when you make the whole sentence (or as much of it as you can grasp) serve as your guide in choosing glosses or grammar formulas, so that you may even find yourself unable to use exactly what you find in the dictionary and grammar and are forced to 'guess in' something that fits. Thus a simultaneous MT and editing is not a logical addition, MT + editing, but instead a logical multiplication, MT x editing, with each of them limiting the other: logically quite a different thing.

Now in my translating from Finnish into English, the control of the process — I am using the word 'control' as a technical term: that which gives warning of blunders — the control of the process was identically semantics. The danger-signal was that the translation didn't make good enough sense. How good is good enough? How can

you, by the semantic control, ensure a good-enough translation? And are there useful implications for pure MT? These are not easy questions. But then, I understand that you don't expect perfect translation either, so that you may well be content with my approximate answers. I shall begin with exact semantic theory, and make the discussion finite by making it approximate.

The pattern of a language, and likewise the linguistic structure of any text, is that network of absolute restrictions (forbidding many or most random occurrences) which is called 'grammar' in the broadest sense of the term. This grammar may be described as the set of impossibilities of combination of linguistic units. On the standard probability scale from zero ('impossibility') to unity ('certainty' or 'strict implication'), we find that grammar uses only the two ends of the scale. For instance, in English it is 'impossible' to have *this* immediately followed by *men;* and the actually occurring open sequences of *this* and *men* 'necessarily, by strict implication' have something between, e.g. a comma or the sequence *group of.* Since impossibility is single, while this implication is a variable (a discrete variable, of which I have cited two values out of an indefinitely long list), it is cheaper to define grammar as a set of impossibilities than to describe it as a set of implications.

Some of the implied items are commoner than others; for instance, *this group of men* is commoner than *this flock of men.* I mean that one of these utterances is more often spoken or printed than the other; and I am not referring to another fact, of a different order, namely that groups of men are commoner than flocks of men — a separate question which might well occupy us elsewhere. We consider now the disparate or 'variable' commonness of the grammatically possible linguistic items, such as *group of, flock of,* etc., in the context *this...men,* and we give this whole phenomenon, proper to this and to all other contexts, a technical name. I call it the 'inside semantics' of the language; and the 'inside meaning' of each item is by definition the statistics of its occurrence in context with other items. Whether there is also something else, independent of this, which could be called the 'outside semantics' *of the language,* is a separate matter to be discussed later; but each 'content' item of a text has, by definition of the word 'content', a proper 'outside meaning' which is simply its referent, the real-world thing-and-event complex to which it refers in this occurrence.

Now the ‘inside meaning’ of each linguistic item in the text or in the language, and that system thereof which I call the ‘inside semantics’ of the language, are indisputable facts. I mean simply that people do not utter all grammatically licit utterances with comparable frequencies (oftennesses); rather, they seldom make such perfectly ‘true’ remarks as “I have never heard a green horse smoke a dozen oranges.” The mathematics of this state of affairs, together with what was previously said about grammar, can be covered by a single statement, thus: ‘The probability of given linguistic items occurring in given linguistic contexts is measured on the standard probability-scale running from zero to unity; the two ends of the scale define the grammar of the language, and the rest of the scale, the open interval between zero and unity, defines the inside semantics of the language.’ The objection has been raised that only a hopelessly long research program could work out the inside semantic system of a language. But that is merely a matter of degree. No statistician claims absolute precision anywhere. From a small body of data, a rough set of statistics will emerge; from a larger body of data, more precise statistics, and so on indefinitely. Impossibility of attaining absolute precision is not a legitimate argument against the existence of the phenomena being investigated. On the other hand, if the statistics do get more precise as the body of data expands, this is customarily taken as a valid argument in favor of the ‘existence’ of what is being statistically treated. Now the G. & C. Merriam Company finds it economically advantageous to expand its citation-file of word-occurrences and contexts indefinitely, while putting far less money into factual reference-books. I assume, therefore, that the existence of ‘inside semantics’ and of ‘inside meanings’ may safely be taken as established.

Let me return briefly, now, to the semantic control of my behavior in translating Finnish into English. When I had applied dictionary and grammar mechanically and thereby manufactured some English nonsense-text, what was the nature of the semantic control that told me it was nonsense? Was it (1) that the English words were strange bedfellows in the sentence; or was it (2) that the English sentence did not match the real world as I knew it? A case could be made out for either explanation, or for a combination of them in any proportion. I am going to try to make out a case for the first theory, the strange-bedfellows explanation.

What causes the hesitation between this and the real-world or outside-meaning explanation? I believe it is the fact that we keep passing knowledge back and forth between two containers, namely (1) language

and (2) sensation-and-manipulation. Very little knowledge, and that only of special kinds, is normally kept, even a little while, entirely inside one of those two containers. The extreme case of knowledge kept entirely inside language is: pure mathematics. The extreme case of knowledge kept entirely inside the other container, namely sensation-and-manipulation, is: the non-language arts, notably painting and sculpture. It is therefore easy to see why neither of those fields is dependent upon national languages for international currency. I have friends who read Russian mathematical publications without knowing a hundred words of Russian; and it is certain that a competent art teacher needs no language in his guidance of a pupil, just as we need none for enjoying a painting. But when mathematicians and artists communicate with others who are not of their own guild, in either case ordinary language always steps in. When pure mathematics is brought to bear upon engineering or atomic physics, ordinary language is always (I think necessarily) used as the mediator; and we all know that painting does get discussed interminably in ordinary language. Thus, as soon as society at large tries to profit from either extreme case, it gets assimilated to the intermediate, majority-party or normal cases. Therefore we can neglect both extremes here, and concentrate on the great majority of human fields of knowledge and action, where the general rule holds: all the socially significant knowledge continually gets passed back and forth between language on the one hand and sensation-and-manipulation on the other hand.

It is in the second container for knowledge, namely sensation-and-manipulation, that we would expect to find the previously mentioned ‘outside semantics’. Although I would not deny its existence in the extreme case of painting or sculpture, or pure music or pure dance, I can nevertheless deny that this is ‘outside semantics of the language’ and proceed inductively to the conclusion that any ‘outside semantics’ as a system, independent of the system called ‘inside semantics’, does not belong to the language at all and therefore does not concern us as linguists or MT workers. Insofar as a semantic system exists which ties all ‘outside meanings’ together into a system, that system is inevitably isomorphic with the system already denominated ‘inside semantics’. The isomorphism is maintained, and fostered as the culture and the inside semantics evolve, by the oscillation of knowledge between language and that non-language realm which I called sensation-and-manipulation. Every new thing, new sensation, or new manipulation, promptly gets named, and the discussion thereof gets standardized, with a standardization faithfully manifested in those contextual occurrence-statistics which I called the ‘inside semantics’

of the culture's language. It is only thus that the innovation can strike root and survive in the culture. Incidentally, this is surely the reason why new schools of painting nowadays change and fade and vanish and get supplanted in such a dizzy dance: deliberately devised to defy discussion in ordinary speech, they die for lack of the cultural survival-value of such discussion. Conversely, any new linguistic coinage which does not get firmly attached to either an old or a new sensation-and-manipulation item is forgotten in a few weeks or years: we call it slang. The exception here is pure mathematics: there the new formulations, devised from the beginning in such terms as make them independent of sensation-and-manipulation, can survive indefinitely and compensate us for the evanescence of slang and of artistic innovations.

With the extremes cleared out of the way, I can now concentrate on the overwhelming majority of human concerns, namely those in which the knowledge does get continually passed back and forth between language and sensation-and-manipulation. And I shall henceforth take it as certain that the 'outside semantics', which I allowed to be possibly autonomous in e.g. the graphic arts, is in these principal human concerns not autonomous at all, but necessarily isomorphic with the 'inside semantics' of ordinary language.

We would seem to have three technical terms left here: 'outside meaning' for one, 'inside meaning' for another, both proper to single linguistic items; and a single 'semantics'. The latter, originally defined as 'inside semantics' and as the system of 'inside meanings', appears now also to be the system of 'outside meanings'. Therefore, each 'outside meaning' is homologous (similarly placed in the system) to the 'inside meaning' of the same linguistic item; and thus we see that the distinction between inside and outside is otiose; they are equivalent and can be treated 'as if' identical, which is the same thing as treating them 'as' identical—even though originally defined quite separately. Thus the fact that windows are made of glass and are breakable and transparent if sufficiently clean—this outside fact does not need to be treated any differently, in linguistic or MT discussion, from the statistical fact concerning English utterances, that the word *window* occurs frequently in context with such words as *glass, broken,* and *wash.* Therefore, all sensation-and-manipulation facts can be built into an MT treatment as soon as they are sufficiently known, and it doesn't matter whether a detail of programming is based upon study of texts or study of the real world—as long as we confine ourselves to the open interval of the probability-scale, between zero and unity probability, for the two ends will govern a separate area of the

programming, namely the grammatical area. Originally I was tempted to claim that semantic programming should be based only upon statistical text-analysis, but now I see that it doesn't matter.

Now semantic programming would seem to be the proper theme of my paper. I have no interest in describing an MT method which is essentially defective in that respect. But since my paper is abstract only, I won't describe the non-defective machine either. I have already done that in another place: Lang. 32.296-7 (1956). Briefly, it is an imaginary machine which uses the presence of each content-word in a sentence as a guide in choosing among the possible renderings of each other word in the sentence, just as a human translator does.

But it may be worth while to add a few words here on 'denotation' and 'connotation'. In a decently written treatise, each content-word of a sentence has a definite reference as its outside meaning. Each occurrence of such a word, that is, means one certain 'thing' (in the broadest sense of 'thing'). But that same word in other occurrences (within or outside of this treatise) might just as well (though either more or less often) have one or more other outside meanings. Now the outside meaning which the word has in the sentence in question is its 'denotation' there. This denotation, being particular, is exempt from perturbation by the rest of the sentence. On the other hand, all the denotations which that same word could have in still other sentences are its 'connotations' in *this* sentence, to a first approximation. To a first approximation only, however, for these connotations (unlike its denotation here) are very much subject to perturbations from the denotations and the connotations of the other words in *this* sentence. It is one of the characteristics of skillful writing that this whole perturbation-field is carefully adjusted in two ways: so as to attenuate misleading connotations, and also so as to reinforce helpful connotations. It is not entirely outside the scope of the imagination to make a mental construction of a bilingual machine that would do some of this in MT work. But it is a problem of a different order of complexity than those previously envisaged. You see, the denotations are differently distributed over the vocabularies of the two languages. Therefore, the connotations cannot be simply carried over from S language to T language. Instead, within each sentence at least, the carefully adjusted complete perturbation-field in the S sentence has to be replaced by an adjusted perturbation-field in the T sentence, and this T field has to be composed mostly of connotations differently distributed among the denotative words. It is at present of course a desperate problem; but if the engineers of another millennium solve it, they can do MT of prose literature.