[Paper presented at the first MT conference, June 1952, MIT. Not published.]

MICROSEMANTICS

(Victor A. Oswald, Jr.)

As I have said before, I am persuaded that we must devise the product of MT in some fashion that will make it, at the conclusion of our process, more or less immediately intelligible to a monolingual specialist in the field to which the translated text pertains. A general pre-editor (monolingual in the FL) can perform all sorts of useful tasks; as you will see, I should like to assign him among other duties that of instructing the machine where to find the fracture surfaces of German compounds. A general post-editor (monolingual in the TL) should be required to shape the ends the machine has rough-hewn, in particular to smooth out the word order where the machine may have jumbled it. Above all, I should want to assign to a post-editor the task of choosing the most satisfactory preposition from the battery we shall have to provide in the TL for each preposition in the FL, and likewise to choose the appropriate equivalent from among the less formidable array of multiple choices for conjunctions. A monolingual general pre-editor, however, cannot efficiently aid in the interpretation of meaning-bearing words that have diverse significance in diverse contexts. A monolingual post-editor will inevitably be confronted with all sorts of contexts with which he is not familiar. Bilinguals, as I have pointed out, will be only a little better equipped to cope with the problem. For, if we assume we can resolve patterns of syntactic connection from the FL into the TL — as I am sure we can — and if we assume that we can supply at a reasonable speed all the possible significances in the TL for each meaning-bearing word in the FL, then the competence required for the ultimate interpretation of the text, is not linguistic insight or skill, but Specialized knowledge of the field in question. Our process would then require an FL pre-editor, a machine, a TL post-editor and a rather large battery of TL experts.

I am of the persuasion that this process would produce satisfactory mass translations. Whether it would be economical or not, someone else will have to decide.

Hypothetically, however, an alternative arrangement is possible: to replace the battery of specialists by a series of permanent micro-glossaries, each of which would provide no more than two-to-one, and a preponderance of one-to-one, TL equivalents for the meaning-bearing words in this particular FL context. Let me introduce you to the micro-semantic hypothesis.

It is now well known that the data obtained from every sort of linguistic frequency count fall into the graphic pattern of a parabolic curve. Words of highest frequency drop in an abrupt descent, words of medium frequency curve out slowly, and the graph line ends on a long and dismal flat – a line extending presumably to infinity, composed of words that occur once. The upper segment of the line contains words which will dispose of the vast majority of all running words in a context – usually in the neighbourhood of eighty percent. But these, alas, are chiefly the little “thes”, “ands”, and “buts”; while the meaning-bearing forms, particularly the nouns, are out on the tail of the curve. Thus, to get at the meaning-bearing forms of a specific context, we must, so to speak, put off and isolate a specific segment of the parabolic curve. Happily, however, as you will recall, the data of <u>all</u> frequency counts fall into the same pattern, which means that a frequency count of any micro-segment of any language – say the nouns in German contexts pertaining to brain surgery — should give a parabolic curve whose high-frequency elements ought to dispose of eighty-per-cent of all running nouns. A corollary to this hypothesis would state that eighty per-cent of all the running nouns ought to convey the bulk of the semantic load borne by the nouns in the context in question.

My present assistant, Mr. Richard Lawson, and I set out fairly recently to test this hypothesis. Although our work is not finished, I can report to you that there can be no doubt the hypothesis and its corollary are true.

We began, no doubt too modestly, by testing this hypothesis only on the technical noun vocabulary of brain surgery. We abstracted the technical nouns from a first German article on brain surgery, tried them out on a second, added the technical nouns the second article, tried out the new total glossary on a third article, and so on up to sixteen articles in all, amounting to about two hundred pages of texts. Each succeeding article was chosen from a different field of brain surgery. Our goal is an optimum glossary of technical brain surgery terms, and we cannot be sure that we have obtained it until we have done much further testing. The amazing fact, however, is that from the outset the glossary took care of more than 80% of the technical running nouns and from the fourth article onward covered an average of 80% of all the technical noun <u>items</u> in each succeeding article. Moreover, as I shall show you, the semantic level is so provided for that the text is usually intelligible — technical-noun-wise, that is — in spite of the fact that about 15% of the running nouns (about 20% of the items) are not translated.

What is possibly more interesting was our discovery that a similar glossary non-technical nouns can be compiled. We soon became aware that the non-technical nouns, which we had not been including in our glossary, kept reappearing from article to article. We finally retraced our steps and compiled a non-technical glossary to parallel the technical glossary, and we ultimately found that the frequency of its forms was structured analogously to that of the technical. I do not mean, of course, that we were surprised to find the parabolic curve again. What I mean is that the high-frequency elements of this list recur as do the high-frequency elements of the technical noun list, and the gaps left in the context by nouns whose occurrence is not predicted do not obliterate the essential meaning of the passage (noun-wise only, please recall). The appendix supplies samples of the efficiency of these glossaries.

In other words, brain surgeons writing on brain surgery are not only compelled to choose their technical nouns from a limited vocabulary, but their patterns of communication are so limited by practice and convention that even the range of non-technical nouns is predictable.

As I have told you, our work is by no means finished. We must determine whether or not the efficiency of the glossary can be increased by further additions, although I anticipate that eighty per-cent of all running nouns and forms will prove to be the highest average we can attain when we apply the glossary to new contexts. On the other hand, when the glossary is applied to fields of brain surgery already examined, the average appears to be consistently above 85%: some spot checks give an average of between 95% and 100%. It may be, of course, that our field is closer to a closed system than we have assumed, and that we can ultimately obtain better results, at least for the purely technical noun vocabulary. Above all, we shall want to check our non-technical vocabulary on other fields of medicine. Possibly the non-technical glossary may prove to have a 'carry over' value. Spot checks in other fields of surgery have been encouraging.

Perhaps the optimum system for our medium will prove to be a generalized non-technical glossary for the whole domain of medicine, along with specialized glossaries for fields of specialization.

Finally, we should like to know how a general non-technical glossary might operate for scientific fields related to medicine. And – alas – we know what we now know only about nouns. Verbs, adjectives, and possibly adverbs should be similarly investigated.

Frankly, I do not know today just how significant our findings are for MT. They indubitably indicate that micro-glossaries can be constructed that could serve to replace a team of specialists in our proposed process of MT. Their ultimate efficiency remains untested, however, and it is possible that it might be prohibitively expensive to produce them.

APPENDIX

A word about the structure of the micro-glossary. The German system of noun-compounding is such that a glossary based on the graphic forms would be both unwieldy and grossly inefficient because of unnecessary repetition. Almost any sequence of nouns in German not syntactically connected is automatically made into a compound, and your German nouns strays gaily about appearing now as the "head" and now as the "tail" of a compound. In a passage at my elbow as I write this I see Hirnödem, Hirnprolaps, Hirndrucksteigerung; and also Liquordrucksteigerung, Schädeldruck, Liquordruck, Druck, Ödem, Prolaps, Steigerung, Hirn, Schädel. Twelve forms, many of them formidably long. But if you abstract Hirn, Ödem, Prolaps, Druck, Steigerung, Liquor, you can do the same work with six forms. In a word, you must break up German compounds if you want to make any sort of efficient German-English glossary. And so we have simply humanized the "awful German language" by dissecting its compounds when they occur, into their components. We know no mechanized process by which this could be accomplished, but an intelligent FL pre-editor could indicate the dissection for any sort of context.

Please remember that the purpose of the samples below is to show how much of the semantic load, noun-wise, is covered by the glossary. Above all, remember that the translations assume the solution of all problems except transference of noun-meaning.

SAMPLE A. A critical passage from Article XVI (on which we scored only 87% of technical running nouns, 85% of all technical items, 90% of non-technical running nouns, 84% of all non-technical items: the respective figures for Article XV were 99%, 96%, 90%, 84%.)
From "Über einem operierten Fall eines verknöcherten subduralen Hirntons traumatischen Ursprungs" (Dr. Emmerich Lang), Zentralblatt für Neurochirurgie, VII (Nr. 5/6), 1952, 193-202.

(Nouns from the technical glossary are completely underlined; those from the non-technical glossary are marked by broken underlining [here *italic*]. Fracture points of compounds are indicated by a diagonal. Words not in our glossary are supplied in the footnotes.)

Aus den vorstehenden, auf unserem *Fall* bezüglichen genauen *Angaben* müssen wir gewisse *Punkte* hervorheben und zum *Gegenstand* einer *Untersuchung* machen.

Bei der ([1]) trat im 23. *Lebens/jahre*, 3 Tage nach einem die linke *Seite* des *Kopfes* treffenden Trauma die Epilepsis auf, ohne schwere Lähmungs/*erscheinungen*. Die auf der rechten *Seite* des Schädels, gerade in der dem Krankheits/*prozess* entsprechenden Region wahrnehmbare teilweise *Unregelmässigkeit* und abnorme Knochen/*dicke* sind vielleicht kein zufälliges ([2]), und dieser *Umstand* kann auf einem älteren *Ursprung* des intrakraniellen Krankheits/*prozesses* hinweisen (*Geburts*/trauma?). Auf welcher *Zeit* immer wir jedoch den *Prozess* zurückführen, die *Rolle* des Traumas erschient zweifellos. Ebenso müssen wir annehmen, dass der umschriebene Blut/erguss zwischen Dura und Hirn/rinde verkalkte und verköcherte, weil sowohl ein Entzündungs- oder Nekrose/*prozess*, wie auch ein Tuberkel, die ([3]) eines parasitären *Prozesses* oder die *Möglichkeit* der ([3]) einer Geschwulst ausgeschlossen werden können. Zu welcher *Zeit* die ([3]) und Verknöcherung erfolgten, konnte weder durch das klinische *Bild* noch durch die *Untersuchung* mit *Sicherheit* festgestellt werden.

TRANSLATION OF SAMPLE A:

From the exact *data* above pertinent to our *case* we must call-special-attention-to certain *points* and make (them) the *subject* of-an *investigation*/*examination*.

In-the-case-of this ([4]) epilepsy occurred in her 23rd *life-year*, 3 *days* after a trauma striking the left *side* of-the *head*, without grave paralysis-*phenomena*. The partial *irregularity* and abnormal bone-*density* perceptible on the right side of-the skull directly in the region corresponding-to the disease-*process* are perhaps not an accidental ([5]), and this *circumstance* may point to an older *origin* of-the intracranial disease-*process* (*birth*-trauma?). To whatever *time* we trace-back the *process*, however, the *rôle* of the trauma appears indubitable. Likewise we must assume that the circumscribed blood-*effusion* between dura and brain/cranium-cortex calcified and ossified because both an inflammation-

---

[1] Kranken. Adjectives used as nouns were not included in our count. They would, of course, appear as nouns in any glossary that included adjectives in general in its scope.

[2] Zusammentreffen. Verb used as noun. Cf. Remarks in footnote 1.

[3] Verkalkung. This word, also, occurred 16 times in this article, the technical noun of highest (!) frequency, and the only one of frequency higher than 5 not previously contained in our glossary. For this reason I have chosen a passage in which it occurred.

[4] Patient.

[5] Coincidence.

or necrosis-*process* as well as a tubercle, the ([6]) of a parasitic *process* or the *possibility* of the ([6]) of a tumor can be ruled-out. At what *time* the ([6]) and ossification took place could be established with *certainty* neither by the clinical *picture* nor by the *investigation/examination*.

SAMPLE B: A spot check of a new article: percentage of coverage not calculated.

From "Über motorische Reizerscheinungen bei peripheren Nervenverletzungen" (Dr. W. Goetz, Dr. H. Becker), Zentralblatt für Neurochirurgie VII (Nr. 1-5), 1943, 129-142.

Den *Ausfalls/erscheinungen* nach Läsion peripherer Nerven stehen die Reiz/*erscheinungen* gegenüber. Während erstere zusammen und isoliert vorkommen (d.h. neben der gemeinsamen *Schädigung* der motorischen und sensiblen *Funktion* können allein die sensible *Leistung* oder ausschliesslich die motorischen geschädigt sein), werden die Reiz/*erscheinungen* sehr viel häufiger isoliert beobachtet. Die Kenntnis der sensiblen Reiz/*erscheinungen* ist ([7]); es handelt sich dabei um das *Phänomen* der ([8]), das überleitet zur ([9]) als der ausgeprägtesten und eindrucksvollsten *Form*. Demgegenüber ist über motorische Reiz/*erscheinungen*, die bei zentralen *Schädigungen* ganz geläufig sind, bei peripheren Läsionen ausserordentlich wenig bekannt. Und doch stellen sie keineswegs eine *Seltenheit* dar. Sie bieten sich dem ([10]) in dreierlei *Form*:

1. als einmalige *Kontraktion* im *Moment* der *Verletzung*,
2. als tonische *Kontraktur* des Muskels bald nach der *Verletzung* oder später,
3. als rhythmische *Bewegung* der Muskulatur ohne (fibrilläre, fassikuläre Zuckung) oder mit lokomotorischem Effekt (Kloni, Tremores, athetoide *Bewegungen*).

Nicht hierher gehört die myogene *Kontraktur*, die ausschliesslich durch den *Funktions/mangel* gelähmter Muskel/*anteile* im *Verein* mit gewissen prädilektiven *Stellungen* der *Glieder* zustande kommt.

TRANSLATION OF SAMPL B:

To *failure-phenomena* after lesion of peripheral nerves stimulus/irritant-*phenomena* contrast. While the former occur together and isolated (i.e., along-with the general *impairment* of the motor and sensory *function* only the sensory *performance* or exclusively the motor can be impaired), the irritant/stimulus-*phenomena* are observed very much more frequently isolated. The *knowledge* of sensory irritant/stimulus-*phenomena* is ([11]); in-this-connection it-is-a-matter of the *phenomena* of ([12]), which leads to ([13]) as the most-pronounced and most impressive form. On-the-other-hand, extraordinarily little is known in-connection-with peripheral lesions about motor irritant/stimulus-*phenomena*, which are very frequent in central *impairments*. And yet they are by-no-means a *rarity*. They offer-themselves to the ([14]) in threefold *form*.

1. as single *contraction* at-the *moment* of *injury*,
2. as tonic *contraction* of-the muscle soon after the injury or later,
3. as rhythmic movement of the musculature without (fibrillar, fasciculary spasm) or with locomotor *effect* (kloni, tremors, athetoid *movements*).

Myogenous contraction does not belong here, which is-produced exclusively by the function-lack of crippled muscle-*parts* in *union* with certain predilective *positions* of the *limbs*.

---

[6] Calcification.
[7] Allgemeingut.
[8] Hyperpathie.
[9] Kausalgie
[10] Beobachter.
[11] Common property.
[12] Hyperpathy.
[13] Causalgia.
[14] Observer.