MODERN LANGUAGE FORUM

Formerly MODERN LANGUAGE BULLETIN, Established 1915

Volume XXXVIII SEPTEMBER-DECEMBER 1953 Number 3-4

AN IDIOGLOSSARY FOR MECHANICAL TRANSLATION*

To ESTABLISH effective bilingual glossaries for mechanical translation will require some sort of solution for three problems: how to predict the words that will occur in the foreign-language contexts to be translated: how to predict the English meaning of the words in those contexts; and how to reduce to a minimum the searching time required to locate any item in the glossary.

Let us recognize at the outset that these problems have by no means been solved for translation in general. Anyone who has worked at translating, say, from German into English knows that the standard desk dictionaries have at best a limited effectiveness ; that they do not list, for instance, any but the commonest technical terms; and that they seldom explore the semantic range of the forms they list. The larger encyclopedic dictionaries provide richer fare, but make it harder to digest. The number of forms listed is increased, but the searching time is thereby also increased to such a degree that efficiency is markedly reduced. Moreover, even the most encyclopedic of the encyclopedic dictionaries are inadequate guides to any sort of highly technical discourse. Finally, not even the so-called scientific dictionaries, which are devised specifically for the translation of technical material, provide anything like an exhaustive listing of technical terms; and they assume the availability of general dictionaries for. non-technical terms. Thus the well-equipped translator of technical contexts must ideally be provided with both a scientific

*The project of research here reported upon was made possible by a grant from the Rockefeller Foundation, for which we wish to express our gratitude. We owe a further debt of gratitude to Dr. Harry D. Huskey, who arranged to have the grant administered by the Institute for Numerical Analysis.

"Idioglossary" was the term adopted by consensus at the Rockefeller Foundation sponsored Conference on Mechanical Translation (Massachusetts institute of Technology, June 17-22, 1952) to designate any word-list devised for the translation of material in a rigidly limited field of specialization.

and an encyclopedic dictionary, must be prepared to squander an incalculable amount of searching time, and must ultimately do his own research or consult a specialist in the field in question to determine the meaning of forms not listed in the dictionaries at his disposal.

Actually, of course, the most effective translator of technical material has the bulk of the necessary vocabulary at his fingers' tips, and needs to have recourse to dictionaries, or to seek consulation, only for a relatively small number of forms. Let us admit that any foreseeable apparatus for mechanical translation can only hope to emulate such a translator's facility. If we can devise special glossaries to be scanned by electronic devices, then a mechanical operation might compete with, or even surpass, a translator's more or less instantaneous translation of terms with which he is familiar; but no mechanical device can possibly cope with the problem of unpredicted forms. The coinage of new words and of new meanings for familiar words, which is constantly practiced—and must be practiced—by writers of scientific discourse, will alone make perfect predictability impossible; and even a highly efficient prediction of established terms will always leave an unpredictable balance. Thus the reader of a text produced by mechanical translation will find himself in the position of the translator who has a residue of unfamiliar forms that he can interpret only by having recourse to a dictionary or by consultation with an expert. Since, however, it is at present the intention to produce mechanically translated texts only for experts in fields of specialization, it might well be that such an expert could extract the meaning from a text in spite of the residue of forms left untranslated.

Among the linguisticians who are investigating the feasibility of mechanical translation there are two schools of thought. One group believes that investigation should be directed toward finding theoretical formulations on the level of general language: that is to, say, to provide theoretical operations which would mechanically translate *in toto* any context from a given foreign language into English. It is the contention of this group that, once such theoretical operations have been devised, technicians could then design a machine to perform the operations. Many of the formulations evolved by advocates of the theoretical approach

have been admirable, some have been brilliant; but they are as yet only signposts at the head of a very long road.[1]

The other school is made up of pragmaticians, who hold that, if we limit our operations for the time being to scientific discourse, and if we are content to aim to produce very crude translations presumably intelligible only to experts in a given specialized field, then mechanical translation—using devices now in being or modifications of such devices—might well be realized in the near future. It was the consensus of the pragmatically minded at the Conference on Mechanical Translation that the scheme of syntactic resolutions evolved by Oswald and Fletcher[2] was too elaborate; that a word-by-word translation in the word order of the foreign language context would be intelligible to the expert for whom the text was to be rendered into English; and that what was most urgently needed was to investigate whether or not effective idioglossaries could be produced.

Happily—happily, at least, for those favorably disposed to the success of such an undertaking—it can be demonstrated that, within the accepted limitations, idioglossaries can be devised for effective mechanical translation. For the past year and a half the authors of this report have been testing what we call the idioglottic hypothesis: that it should be possible to apply frequency-count techniques to a microsegment of any language, and to arrive at a predictability of about 78% for the forms within the area selected. Such a microsegment could be, for instance, the technical vocabulary of any tightly restricted field of specialization, or any subdivision of such a vocabulary such as the nouns alone or the verbs alone. Let us elaborate the point before we proceed.

It is now well known that the data obtained from every sort of linguistic frequency count fall into the graphic pattern of a descending monotonic curve.[3] Words of highest frequency drop in an abrupt descent, words of medium frequency curve out

[1] Cf. especially Erwin Reifler's eight papers in the sequence: *Studies in Mechanical Translation* (dittograph script), and Y. Bar-Hillel's "A Quasi-arithmetical Notation for Syntactic Description," *Language,* Vol. XXIX, No. 1 (Jan.-Mar., 1953), pp. 47-58.

[2] Victor A. Oswald, Jr. and Stuart L. Fletcher, Jr., "Proposals for the Mechanical Resolution of German Syntax Patterns," *Modern Language Forum,* Vol. XXXVI. No. 3-4. (Sept.-Dec.. 1951), pp. 81-104.

[3] Cf. George K. Zipf, *Human Behavior and the Principle of Least Effort* (Cambridge, Mass.: Addison Wesley Press. 1949), and William E. Bull, "Natural Frequency and Word Counts," *Classical Journal,* Vol. 64, No. 8 (May, 1949), pp. 469 ff.

slowly and break into "steps," and the graph line ends on a long and dismal flat—a line presumably approaching infinity, composed of words that occur only once. The upper segment of the graph line is comprised of forms which will dispose of the vast majority of all running words in the context from which they are abstracted—usually in the neighborhood of eighty percent. But these, alas, in any generalized count, are chiefly the functional . forms, the little "thes", and "ands," and "ofs," while the content-bearing forms, particularly the nouns and the verbs, are well out toward the tail of the line. To get at the distribution of the content-bearing farms we should have to be able to do a separate analysis of the lower part of the line.

In general language contexts, however, there is never a sufficient representation of the vast potential stock of forms,[4] with the result that an analysis of content-bearing forms in general language for purposes of prediction is utterly impractical. We reasoned, however, that the distribution of content-bearing forms in a microsegment of the language might well be different; for what else could be implied by the familiar fact that every branch of every science has its own special jargon? Since we knew that any frequency count would fall into the familiar pattern, it might well be that an analysis of the technical components of, say, the nouns in German contexts pertaining to brain surgery, would produce a curve whose high-frequency elements could be expected to predict 80% of all technical running nouns in such contexts. A corollary to this hypothesis would state that eighty percent of all the technical running nouns ought to convey the bulk of the semantic load borne by the technical nouns in the context in question. Moreover, in any specific technical discourse the technical nouns are likely to have only one referent, with the result that they either will have, or can be assigned, one sole analogous referent in another language. (German *Rinde* in brain surgery contexts is equivalent only to English "cortex"—never to any of the other possible referents such as "bark," "rind," or "crust." The idioglottic hypothesis implied, then, that both the problem of predictability of forms and the problem of predictability of mean-

[4] There are. for instance, in James Joyce's *Ulysses*—a novel famous for its author's yirtuoso performance with the English vocabulary—only 29,899 separate *forms,* including conjugated forms of verbs, plurals of nouns, inflected forms of pronouns, etc. (cf. Miles Hanley, *Word Index to James Joyce's Ulysses,* Madison, Wis., 1937). This figure clearly represents only a trivial fraction of the 550,000 entries in *Webster's New International Dictionary.*

ing ought to be susceptible of solution. What we had no means of estimating was how the hypothesis would apply to the problem of holding searching time to a minimum. It only seemed likely that the number of technical nouns required for an effective glossary might be relatively few.

We began by testing the hypothesis only on the technical noun vocabulary of brain surgery contexts. We abstracted the technical nouns from a first German article on brain surgery, tried them out on a second, added the technical nouns from 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 text. Each succeeding article was chosen from a different field of brain surgery. By the time we had completed the abstraction from the sixteenth article our figure« made it clear that we had reached the saturation point. (In fact, after we had finished all our work, it became apparent that additions from the fourteenth article on had not essentially changed the structure of the material.) But long before then, from the seventh article onward, the glossary had been predicting consistently more than 85% of the technical running nouns, and the figures rose to a peak of 99%. Even more gratifying was the fact that from the tenth article onward the glossary predicted consistently more than 80% of the technical noun *items* (single lexicographical entries), with a peak figure of 96%. Finally, what was most gratifying and most significant of all, was our discovery that a glossary of non-technical nouns could similarly 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 found in time that the frequency of its forms was structured analogously to that of the technical, so that we ultimately entered all noun forms in a common noun glossary.

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

The results of the study of the noun vocabulary clearly

postulated a similar analysis of the two other types of content-bearing forms; the verbs and the adjective-adverbs (the uninflected form of the adjective in German functions as the comparable adverb). The results of the extension of our study far surpassed our most sanguine prognoses. We had expected, in compiling the lists of verbs and adjective-adverbs, to find a much lower level of predictability than we had found in the case of the nouns. Apparently, however, the same psychological "set" that causes brain surgeons to limit their noun vocabulary operates to limit their choice of verbs and adjective-adverbs. The following table, which shows the percentage of predictability of all content-bearing forms from article to article, will indicate how closely alike is the patterning for all three types of forms.[5]

Percentage of Predictability

ARTICLE	NOUNS		VERBS		ADJECTIVES	
	items	*running*	*items*	*running*	*items*	*running*
I	-------	-------	------	------	------	------
II	47%	80%	28%	63%	33%	55%
III	57%	73%	59%	79%	54%	65%
IV	78%	86%	76%	85%	67%	76%
V	69%	88%	75%	88%	70%	79%
VI	70%	92%	74%	90%	69%	81%
VII	63%	78%	69%	80%	60%	80%
VIII	77%.	88%	77% .	88%	76%	86%
IX	75%	85%	74%	89%	72%	80%
X	80%	92%	68%	86%	67%	83%
XI	76%	85%	78%	90%	74%	83%
XII	78%	89%	80%	90%	73%	84%
XIII	81%	90%	85%	92%	77%	85%
XIV	86%	92%	82%	91%	78%	82%
XV	88%	95%	79%	89%	79%	92%
XVI	85%	88%	82%	92%	73% .	85%

5 The initial disparity in favor of the nouns is caused by the fact that the first passage in our series is the brain-surgery section from a handbook on surgery (V. Orator, *Spezielle Chirurgie*, 16. und 17. Auflage, Leipzig, 1942), which presents its material in a “telegram style” that is superabundant in nouns.

As the reader can see, the predictability for all three types of content-bearing forms is ultimately in the neighborhood of 80% for items, and close to 90% for running words. We have further tested the idioglossary by using it for spot-checks of fifteen articles on brain surgery (about twenty pages checked out of a total of 150), and we have found the glossary to be consistently effective within the range indicated by the percentage figures for the last four articles on the table above. To that extent—and, of course, within the accepted limitations—the problems of predictability for mechanical translation appear to be capable of solution.

As for the problem of minimal searching time, since our entire idioglossary consists of exactly 1925 noun forms, 1177 verbs, and 1226 adjectives, for a total of only 4328 entries, it is obvious that the time required to locate any single item is minimized to the ultimate degree. If the glossary could be entered on one or more magnetized drums to be scanned by a high-speed electronic computer, the average search time would be one-quarter of a second per item—which would definitely surpass human speed. Of course, if the glossary could somehow be made available for direct search by a high-speed computer, the operation would be vastly more rapid than a human search.

It is our opinion as linguisticians that the major linguistic problems of pragmatic mechanical translation can now be regarded as solved. We can provide glossaries with an adequate degree of prediction and a minimum requirement of searching time. We can, if it be found desirable, supply routines for the mechanical resolution of syntax patterns. The practical application of our solutions lies, however, not in the realm of linguistics, but in that of technology—or, perhaps, in some unexplored borderland where the domains of linguistics and technology meet. What is now needed is a project of joint linguistic and technological research which would grapple with problems not susceptible of solution by linguistic techniques alone. For example: Can the lexicographical entries be reduced to a code, so as both to limit the size of the individual entry and to facilitate the scanning of the total glossary? How many sub-entries—or, rather, how few—will be required to provide for the declined forms of German nouns and adjectives and the conjugated forms of German verbs? Is it really advisable to abandon syntactic resolution altogether ? Do we not need some resolution to determine the English equivalent of forms functionally ambiguous? Finally, a very elaborate investigation

should be made to determine whether such crude translations as we propose will actually be adequately intelligible to the average practitioner of a field of specialization. When answers have been provided for these questions it will be time to inquire whether the cost of mechanical translation will or will not be such as to make the operation economically practical.

At the present moment we know only that it can no longer be said that mechanical translation is impossible. At present, to paraphrase a fellow lexicographer, it cannot be done well, but it is surely remarkable that it can be done at all.

Appendix

I THE CONSTITUENTS OF THE IDIOGLOSSARY

The verb and adjective-adverb entries are treated as forms not capable of further analysis; *e.* g., verbs that consist of a stem and a compounding prefix *(ab-brechen, ab-dichten, ab-fassen,* etc.) are carried as single entries, for the reason that analysis of such forms seldom provides a recognizable English equivalent. Noun compounds, on the other hand, have been broken down into their constituent elements, and the elements make up the bulk of the entries; e. g., *Hirnstamm* is broken down into *Hirn* ("brain") and *Stamm* ("stem"), and the search is intended to produce "brain stem" as the English equivalent of the German form. Noun compounds whose English equivalents are not recognizable from a juxtaposition of components are listed as compounds; e. g., *Tatsache,* with the English equivalent "fact." Adjective-noun compounds not satisfactorily analyzable are listed as compound nouns; e. g., *Grosshirn* ("cerebrum").

II A GENERAL GLOSSARY OF ADVERBS AND FORMS PRIMARILY FUNCTIONAL

This glossary would, like the idioglossary, be comprised of several main divisions: adverbs of quantity and degree (indeclinable); cardinal numbers up to twenty (indeclinable) : particles (indeclinable) ; personal, demonstrative, relative interrogative pronouns (all declined forms supplied) ; conjunctions; prepositions. These forms are what might be called lexicographical constants, for the reason that they occur with maximal frequency in any and all language contexts. The total of all such forms, including the oblique-case representatives of those that are declinable, would run to about 1,000 entries. A glossary of these forms would have to be used in conjunction with an idioglossary.

III A HYPOTHETICAL OPERATION OF MECHANICAL TRANSLATION

As in the hypothetical operation outlined in the "Proposals for the Mechanical Resolution of German Syntax Patterns" we shall suppose that a German text—this time a brain surgery article—is supplied from an electric typewriter to a high-speed electronic computer. All forms are introduced in one font of letters, i. e.. capitalization is abandoned. Punctuation is rigidly retained as in the original text.

Each form supplied (identifiable as a separate form by ordinary word spacing) is compared by the computer with a register of forms available on one or more magnetized drums connected to the computer. This register would be comprised of the general glossary and our particular idioglossary. Every German' item in the register would be provided with one or more English equivalents,

which would be supplied to an output typewriter. One equivalent, at most two, would be sufficient for the items of the idioglossary. Most of the items of the general glossary, excepting only the adverbs, numbers, and particles, will require either multiple equivalents or will have to be subjected to syntactical analysis. In the event that syntactic analysis is not abandoned, each German item in the register will have to carry some marker to indicate its syntactic function. Routines for syntactical analysis would take up all of the capacity of the computer not needed for the operation of comparing supplied forms with those in the register.

Our operation could now proceed in one of two ways. Without syntactical analysis each German form in sequence would be represented in the translation by its English equivalent or equivalents. The word order would be that of the original text. The choice among multiple equivalents would be left to the reader, as would the resolution of the German word order. Although this system has much to commend it, we shall not reproduce here the text that it would provide; first of all, because it is extremely cumbersome to print a column of multiple equivalents for every form that would require them: second, because those interested can find a specimen of such a text in Yehoshua Bar-Hillel's "The Present State of Research on Mechanical Translation." *American Documentation*, Vol. II. No. 4 (October. 1950). p. 227; third, because we believe that, even though some of the apparatus of syntactic analysis should be discarded (all that pertains to rearrangement of word order), the analysis of functional forms should be retained because of its enormous reduction of the requirement of multiple choice.

Let us now assume that the entries come through in German word order. but that syntactical analysis is applied to functional forms, as outlined in the "Proposals." Specimen texts from a brain-surgery journal would be translated as follows by our process (forms not contained in the idioglossary will appear untranslated in the English text). These texts are a representative sampling whose extremes show the process working at its best (Specimen I) and at its worst (Specimen III). It must be remembered that our percentile figures apply to complete articles or reviews. When, within any article the forms not contained in the idioglossary are scattered, the lacunae that occur in any given paragraph are almost negligible: when the unpredicted forms are sporadically concentrated, the lacunae in the area of concentration considerably impair the intelligibility of that portion of the article, although possibly not to such a degree that the passage would be totally incomprehensible to the specialist.

Specimen I

From: D. Krüger. "Zur Versorgung von Verletzungen im Bereich der vorderen Schädelbasis." *Zentralblatt für Neurochirurgie. 7.* Jahrgang (1942). No. 5/6. pp. 21 1-212.

Die Verletzungen im Bereich der vorderen Schädelbasis stellen den Chirurgen immer wieder vor die Frage, in welcher Form derartige Verletzungen wegen der möglichen Beteiligung der Nasennebenhöhlen (NNH.) am besten zu versorgen sind. Hierbei spielt die Drainage des NNH.-Gebietes eine wichtige Rolle.

Die häufige Miteröffnung sowohl der NNH, als auch der Ohrhöhlen und die damit gegebene Infektionsgefährduag des Schädelinnern hat Tönnis veranlasst, die Schädelbasiswunden von den Wunden an der Konvexität zu trennen. Da bei Schädelbasisverletzungen infolge direkter Infektion, d.h. also von der Wunde ausgehend, mit einer frühzeitig einsetzenden Meningitis zu rechnen ist (siehe

Tönnis, Dtsch.Mil.arzt 1942,4), sind hier also besondere Behandlungsmassnahmen erforderlich.

The injuries in the region of-the anterior cranium base put the surgeon always again before the question, in which form such injuries on-account-of the possible participation of-the nose sinuses (NNH.) at-the best, to treat are. Here plays the drainage of-the NNH.-area an important role.

The frequent-MITEROEFFNUNG as-well of-the NNH, as also of-the ear cavities and the therewith given infection GEFAEHRDUNG of-the cranium interior has TOENNIS caused, the cranium base wounds from the wounds at the convexity to separate. Since in cranium base injuries as-a-result-of direct infection, i.e. therefore from the wound proceeding, with an early setting-in-meningitis to reckon is (see TOENNIS, DTSCH. MIL. ARZT 1942.4), are here therefore special treatment measures requisite.

Specimen II

From: Gerd Peters. "Ueber gedeckte Gehirnverletzungen (Rindenkontusionen) im Tierversuch," *Zentralblatt für Neurochirurgie, 8.* Jahrgang (1943), No. 1-5, p. 175.

Makroskopische Beschreibung der Gehirnveränderungen
a) Gedeckte Gehirnverletzungen

An gedeckten Gehirnverletzungen wurden in den von mir durchgeführten Versuchen epidurale, subdurale, subarachnoideale Blutungen, Kontusionsherde in der Rinde an Stoss- und Gegenstossstellen, Blutungen im Innern des Gehirns und Hirnstamms hervorgerufen. Bei 14% der Tiere wurden keine der eben erwähnten anatomischen Veränderungen gefunden. Da diese Tiere jedoch klinische Erscheinungen zeigten, wenn auch zum Teil in geringerem Masse als diejenigen Tiere, die die eben erwähnten anatomischen Veränderungen aufwiesen, wurden sie in der vorliegenden Arbeit mit verwertet. 5 dieser Tiere starben sogar "spontan". Dieses legt die Annahme nahe, dass wir nach einer traumatischen Läsion des Zentralnervensystems neben schon markroskopisch fassbaren Veränderungen manchmal auch anatomisch nicht nachweisbare Schädigungen setzen, die Spatz als "spurlose Vorgänge" bezeichnet hat, und die gerade bei der Commotio cerebri des Menschen eine Rolle spielen.

Macroscopic Description of-the Brain Changes
a) Concealed Brain Injuries

In concealed brain injuries were in the by me performed experiments epidural, subdural, subarachnoidal hemorrhages, contusion foci in the cortex at STOSS- and against STOSS places, hemorrhages in-the interior of-the brain and brain stem produced. In 14% of-the TIERE were none of-the just mentioned anatomical changes found. Since these TIERE however clinical phenomena showed, if also to-the part in lesser degree than those TIERE, which the just mentioned anatomical changes exhibited, were they in the present work with used. 5 of-these TIERE died even "spontaneously." This lays the assumption close, that we after a traumatic lesion of-the central nerve system alongside-of already macroscopically perceptible changes sometimes also anatomically not demonstrable injuries put, which SPATZ as "SPURLOSE processes" designated has, and which directly in the COMMOTIO CEREBRI of-the human a role play.

Specimen III

From F. J. Irsigler, "Ueber den Heilverlauf experimenteller Hirnwunden bei offener und verlegter Knochenlücke," *Zentralblatt für Neurochirurgie, 7.* Jahrgang (1942), No. 1-3, p. 32.

Für diesen traumatischen Hydrocephalus, dem eine grosse praktische Bedeutung zukommt, kommen mehrere Ursachen in Frage: 1. Eine Verlegung oder Verengerung der abführenden Liquorwege; diese Abflussbehinderung oder -sperre macht sich besonders frühzeitig an den physiologischen Engen der inneren Liquorstrassen bemerkbar, das sind die Foramina Monroi, die Sylvi'sche Wasserleitung und die Ausgänge aus dem IV. Ventrikel. Bei der Verlegung des Foramen Monroi kommt es zum ein- oder doppelseitigen Hyderocephalus der verschlossenen Seitenkammer. Diese Form des traumatischen Hydrocephalus lässt sich auch im Tierversuch beobachten und wurde in unserer früheren Arbeit bereits beschrieben und abgebildet.

For this traumatic hydrocephalus, to which a great practical significance ZUKOMMT, come several causes in question: 1. a displacement or VERENGERUNG of-the ABFUEHRENDEN liquor passages; this ABFLUSS BEHINDERUNG or -SPERRE makes itself especially early in the physiological ENGEN of-the inner liquor channels noticeable, that are the FORAMINA MONROI, the SYLVISCHE water tube and the exits out-of the IV. ventricle. At the displacement of-the FORAMEN MONROI comes it to-the one- or DOPPELSEITIG hydrocephalus of-the blocked side chamber. This form of the traumatic hydrocephalus lets itself also in-the TIER experiment observe and is in our earlier work already described and ABGEBILDET.

Victor A. Oswald Jr.
Richard H. Lawson

University of California, Los Angeles
State College of Washington