Automated Translation of German to English Medical Text

G. WILLIAM MOORE, M.D., Ph.D.
Baltimore, Maryland
U. N. RIEDE, M.D.
Freiburg, Federal Republic of Germany
RICHARD A. POLACSEK, M.D.
ROBERT E. MILLER, M.D.
GROVER M. HUTCHINS, M.D.
Baltimore, Maryland

The feasibility of automated translation of scientific and medical documents remains controversial. This report describes a minicomputer-based German-to-English translation system (TRANSOFT) that employs word order rearrangement followed by word-for-word translation and disambiguation based on context. This translation system was applied to a computer-readable version of Adler's *Knochenkrankheiten (Bone Diseases)*, which contains 118,604 words, with 10,216 distinct words in 7,211 sentences averaging 16.4 words each. The translation required 2,791 word rearrangement formulas, 78 percent of which were first used in the first half of the document. There were 2,392 occurrences of 12 potentially ambiguous terms, of which only 18 (0.8 percent) were not resolvable from the immediate context. As foreign language medical documents become increasingly available in computer-readable form through computerized typesetting, electronic publishing, and improved optical character recognition equipment, automated translation systems may provide a rapid and inexpensive means of obtaining draft translations.

The challenge of automated translation has been studied for almost four decades, but research on practical translators was curtailed in the late 1960s after it was suggested that "fully automated high quality translation" was not economically feasible [1–3]. Subsequent published research in the field has been conducted largely outside the United States or has focused on selected theoretic issues [4–13]. Recently, the Japanese Ministry of International Trade and Industry has launched an ambitious plan for the development of advanced computer technology that includes automated translation between Japanese and major European languages [14]. Automated translation programs are now available commercially [15,16]. With advances in computer technology, automated translation has become increasingly practical for minicomputer and other small computer users. This report describes TRANSOFT, a table-driven German-to-English medical document translation system written in the American National Standard MUMPS programming language, that was used to generate a draft quality translation of Adler's *Knochenkrankheiten (Bone Diseases)* [17,18]. The TRANSOFT system has the advantages of user control of the vocabulary and grammatical rules, portability to a variety of small computers through the MUMPS programming language, and quantitative measures of performance.

METHODS

A recent German language medical text, Adler's *Knochenkrankheiten (Bone Diseases)*, was made available to us in computer-readable form by the publisher, Georg Thieme Verlag [18]. The full manuscript, excluding footnotes, table headings, and figure legends, was written from a nine-track

From the Departments of Pathology and Laboratory Medicine, and the Welch Medical Library of The Johns Hopkins Medical Institutions, Baltimore, Maryland, and the Department of Pathology, Freiburg University School of Medicine, Freiburg i.Br., Federal Republic of Germany. This work was supported by Grant LM-03651 from the National Library of Medicine. Requests for reprints should be addressed to Dr. G. William Moore, Department of Pathology, The Johns Hopkins Hospital, Baltimore, Maryland 21205. Manuscript accepted June 10, 1985.

magnetic tape to American Standard Code for Information Interchange (ASCII) text files on disk on a Digital Equipment Corporation PDP-11/70 minicomputer running Intersystems Corporation's M/11+ operating system and American National Standard MUMPS programming language in the Department of Laboratory Medicine of The Johns Hopkins Medical Institutions. The computer-readable text was subjected to pre-editing and then was translated in its entirety from German to English by TRANSOFT, a sentence-by-sentence translation system written in MUMPS with control information contained in two language-specific translation tables, a word and idiom *lexicon* and a *parsing table* of word rearrangement formulas. The MUMPS programming language (Massachusetts General Hospital MultiProgramming System), which is widely used for medical information processing, was chosen for all TRANSOFT programs because of its powerful character string operators and its string-subscripted arrays with implicit sorting [19–27]. All language-specific control information was incorporated into two external translation tables in order to make TRANSOFT easy to modify and maintain and readily adapted to other translation tasks, such as medical English sublanguages and formal logic [17].

Automated Pre-Editing. Preliminary computer editing and reformating of the raw text file of *Knochenkrankheiten* was carried out to provide a standardized document for the subsequent translation steps. This pre-editing step was performed by special-purpose MUMPS programs, using rules that were applied sequentially to the entire file. All control characters and typesetting commands were first removed. German special characters were rendered in American format, i.e., Ä, Ö, Ü, ä, ö, ü, and ß were rendered as Ae, Oe, Ue, ae, oe, and ss, respectively. Punctuation was reduced to commas, periods, and parentheses as follows. Semi-colons, colons, exclamation points, and question marks were replaced by periods. Asterisks, apostrophes, and quotation marks were replaced by blanks. Each phrase delimited by dashes—such as this one—was enclosed in parentheses. Each hyphenated word, e.g., Bang-Osteomyelitis, was converted to a single non-hyphenated word, e.g., BangOsteomyelitis. Terms containing numeric characters were left unchanged, although decimal numbers were expressed according to American format, e.g., 27.4 rather than 27,4. Punctuation characters other than leading or imbedded decimal points were buffered on either side with a blank (space character) to simplify subsequent steps. The first character of the word at the beginning of each sentence was changed to lower case to simplify the later processing of nouns (which begin with an upper case character in German). With the period as a sentence terminator, each sentence was started on a new line and stored as a separate array element in the text file. This pre-edited text file then served as the source document for all subsequent processing by the TRANSOFT system. Thus the first two sentences of Chapter One:

Das Skelett nimmt im Leben eines Individuums in vieler Hinsicht eine zentrale Stellung ein. Es verleiht jedem Lebewesen seine spezifische Körperform und ist bestimmend für die Architektonik des Körpers.

were modified as follows in the pre-editing step:

das Skelett nimmt im Leben eines Individuums in vieler Hinsicht eine zentrale Stellung ein .

es verleiht jedem Lebewesen seine spezifische Koerperform und ist bestimmend fuer die Architektonik des Koerpers .

Lexicon. A lexicon of words and idioms is one of two external tables of language-specific control information used by the TRANSOFT system. The lexicon consists of all acceptable source language words and idioms, their part of speech designators, and their primary and any alternative definitions. A large portion of the lexicon can be defined initially for a given language pair using published dictionaries, and then augmented with additional vocabulary entries as required for new documents. For our translation, an initial German word list was generated from the pre-edited *Knochenkrankheiten* source file by collating all character strings bounded on either side by a blank, a task easily performed in MUMPS using the $PIECE function. This list was then expanded to include additional noun and adjective declensions and verb conjugations, including separable verb forms. Potential idioms (i.e., multiple word sequences) were added by selecting all occurrences of word pairs, triplets, quadruplets, and so on that appeared at least twice in the source document. Words or idioms were accepted as final lexicon entries by a bilingual speaker, who assigned a *default translation* and a *semantic class* to each entry, using the 19 semantic classes listed below. Most of these semantic classes represent punctuation or ordinary parts of speech, although some reflect the unique requirements of automated translation. For example, U is an ambiguous part of speech commonly encountered in German, and F and Z represent words often encountered in scientific documents that require special processing:

. = period
, = comma
(= left parenthesis
) = right parenthesis
A = adjective or adverb, e.g., aktiv (active), eitrig (purulent)
B = adverb only, e.g., besonders (especially), dadurch (thereby)
C = conjunction, e.g., und (and), aber (but)
D = definite or indefinite article or demonstrative pronoun, e.g., der (the), ein (a), dies (this)
F = foreign word or proper name
H = helping verb, e.g., sein (be), haben (have), werden (become)
I = interrogative or relative pronoun, e.g., welcher (which), warum (why)
N = noun, e.g., Anwendung (application), Auftreten (appearance)
P = preposition, e.g., auf (upon), bei (at)
Q = pronoun, e.g., es (it), sich (itself)
U = verb, gerundive, or participle, e.g., aufgetreten (appeared), entscheidend (decisive)
V = verb only, e.g., auftreten (appear), entscheiden (decide)
Z = number or formula, e.g., eins (one), zwei (two), and

so on, or word containing a numeric character, measurement, or mathematical symbol

[= left bracket (start character)

] = right bracket (stop character)

Each word in the lexicon was also assigned any number of *alternate translations*, which depend upon either the semantic classes of neighboring words or a *context register* of keywords maintained at the time of translation. For example, the word with the greatest number of entries in the lexicon is "der," which has a default meaning of "the" and more than 60 alternate semantic class contexts in which it is translated as "of the," "to the," "which," "whose," or "to which." The final lexicon was then used as the word list for a MUMPS spelling checker program, which examined the pre-edited source file. Nouns at the beginning of sentences that had incorrectly been placed in lower case were detected in this manner.

Parsing Tables. A *parsing table* of word rearrangement instructions, or *parsing formulas*, is the second translation table used by the TRANSOFT system. Parsing formulas are applied recursively by TRANSOFT to transform a sentence in German (source) word order to its corresponding English (target) word order, after which English-to-German word and idiom substitution is performed. These parsing formulas are akin to "scripts," "frames," or "patterns" used in other automated translation systems [6–9,12,13]. An unparsed or incompletely parsed sentence in the source language can be represented by the consecutive sequence of semantic class designators, called a *parsandum*, for that sentence. A parsing formula consists of a *key* (i.e., sequence of semantic class designators to be recognized in the parsandum) and rearrangement instructions for the substring of the parsandum corresponding to the key. For example, the following German sentence:

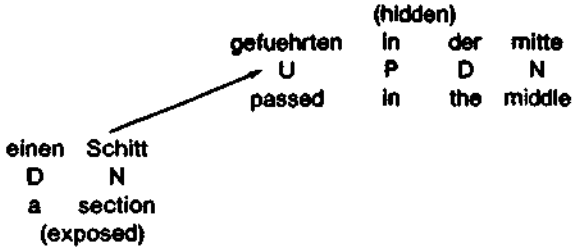
Der	Koerper	laesst	sich	durch	einen	in	der	Mitte	gefuehrten	Schnitt	zerlegen
[D	N	H	Q	P	D	P	D	N	U	N	V]
The	body	allows	itself	through	a	in	the	middle	passed	section	divide

has a parsandum of [DNHQPDPDNUNV], where [and] are the start and stop character delimiters, respectively. The substring DPDNUN is the key to an entry in the parsing table. The rearrangement instructions for this key can be expressed by an arrow diagram:

(old parsandum substring)

(new parsandum substring)

or more compactly by the formula $^{1}D\,^{4}\square P\,^{5}\square D\,^{6}\square N\,^{3}\square N$, where the numeric prescript (preceding superscript) gives the position of the semantic class designator of the new parsandum, and the alphabetic prescript (including the box, □) gives the value of the semantic class designator for that word in the new parsandum. This parsing formula notation can be made even more compact by placing the prescript on the same line as the key and by eliminating the alphabetic prescript when it is the same as its corresponding key element. A box (□) indicates that the word is *hidden* in subsequent recursive steps. In this example, the revised phrase is:



where "gefuehrten in der Mitte" is hidden in subsequent parsing steps. As a general rule, the *hidden* part of the phrase is placed in *target language* (*English*) *word order*, whereas the *exposed* part of the phrase is retained in *source language* (*German*) *word order*.

During recursive processing of each sentence by the TRANSOFT program, either the entire parsandum, or its longest matching substring, is matched to a key in the parsing table, and a reduction is performed. This process is continued until the parsandum contains only one word (parsing complete) or no matching key can be found in the parsing table (error condition). The recursive algorithm is mathematically guaranteed not to cycle indefinitely if every parsing formula contains a box [17]. In our study, sufficient parsing formulas were included to handle every sentence. For our example, there were three parsing steps, as follows:

Parsandum	Parsing Formula
[DNHQPDPDNUNV]	1D4□P5□D6□N3□U2N
[DNHQPDNV]	1Q2□P3□D4□N5V
[DNHQV]	1[2□D3□N4VH6□Q5□V7]
[V]	(complete)

The final, English word order is then determined as:

Der	Koerper	laesst	zerlegen	sich	durch	einen	Schnitt	gefuehrten	in	der	Mitte
[D	N	H	V	Q	P	D	N	U	P	D	N]
The	body	allows	divide	itself	through	a	section	passed	in	the	middle.

The parsing table for this German-to-English translation was generated incrementally by having TRANSOFT repeatedly translate portions of the source document, with a bilingual speaker reviewing successive translations and entering additional parsing formulas. Initially, the empty parsing table caused TRANSOFT to leave the source word order unchanged, with English words simply substituted for the German. This primitive translation then suggested required word rearrangement rules, and the appropriate parsing formulas were entered into the computer interactively. Portions of the source document were then retranslated using the updated parsing table, and the resulting translation was inspected for additional, suggested parsing formulas. This process was repeated until a satisfactory translation was obtained. Using this parsing formula concept, a moderately experienced bilingual speaker can rapidly assemble a parsing table for the types of documents he or she customarily translates.

German-to-English Document Translation. The actual

TABLE I Sample Translations from *Knochenkrankheiten* (pages 1, 168, 318 [18])

Das Skelett nimmt im Leben eines Individuums in vieler Hinsicht eine zentrale Stellung ein. Es verleiht jedem Lebewesen seine spezifische Koerperform und ist bestimmend fuer die Architektonik des Koerpers. Gleichzeitig wird die Groesse des Individuums entscheidend durch das Skelett gepraegt. Diese Formgebung durch das Skelett ist proportional und symmetrisch angelegt, wobei die Groesse der einzelnen Skeletteile dem Gesamtskelett angeglichen ist.	The skeleton comprises in the life of a individual in many regard a central placement. It lends to every life form its specific body form and is determining for the architectonics of the body. Simultaneous the size of the individual becomes impressed upon decisive through the skeleton. This giving of form through the skeleton is laid out proportional and symmetric, whereby is adjusted the size the individual skeletal parts to the skeleton as a whole.
Osteochondrom. Osteochondrome sind die weitaus haeufigsten gutartigen Knochengeschwuelste. Unter den benignen Knochentumoren haben sie einen Anteil von 40%. Es handelt sich um eine knoecherne Neubildung, die von einer breiten Kappe aus hyalinem Knorpelgewebe ueberzogen ist und sich von der Knochenoberflaeche pilzfoermig in die umgebenden Weichteile vorwoelbt. Da diese Tumoren nur langsam an Groesse zunehmen, machen sie oft erst spaet durch eine Schwellung auf sich aufmerksam.	Osteochondroma. Osteochondromas are the far and away most frequent benign bone tumors. Under the benign bone tumors have they a proportion from 40%. One is dealing with a new bone formation, which is covered from a wide cap out of hyalin cartilage tissue and vaults in front of itself from the bone surface mushroom-shaped in the surrounding soft tissue. Since this tumors increase only slow on size, it make aware upon itself often first late through a swelling.
Traumatische Meniskuslaesion. Hierbei handelt es sich um eine mechanische Zerreissung eines gesunden Meniskus meist infolge eines sportlichen oder beruflichen Unfalles. Von dieser haeufigen Verletzung ist vor allem der mediale Meniskus betroffen (zehnmal haeufiger als der laterale Meniskus). Man unterscheidet einen vollstaendigen oder teilweisen Meniskusriss von einem haeufigeren Substanzriss, bei dem der Meniskus gespalten wird. Die Verletzung erfolgt bei ploetzlicher Streckung und gleichzeitiger Rotation im Kniegelenk.	Traumatic meniscus lesion. Hereat one is dealing with a mechanical tearing of a healthy meniscus most as a result of a sportly or professional occupational of accident. From this frequent injury is involved above all the medical meniscus (ten times more frequent than the lateral meniscus). One distinguishes a complete or partial meniscal tear from a more frequent substance tear, at which is cleft the meniscus. The injury results at sudden stretching and simultaneous rotation in the knee joint.

document translation step is straightforward, since all language-specific control information resides in the lexicon and parsing tables just described. The MUMPS TRANSOFT program processes a source document sentence by sentence, through the following steps. The sentence is read from the pre-edited source document disk file into a local array in processor memory. The sentence is $PIECEd on spaces, each word or idiom (i.e., consecutive string of words with an entry in the lexicon) is assigned a semantic class designator, and the initial parsandum for the sentence is generated. Then at each parsing iteration, either the entire parsandum or its longest matching substring is matched to a key in the parsing table, and a reduction is performed according to the word rearrangement instructions specified by the parsing formula. This process is continued until parsing is complete (the parsandum has been reduced to a single semantic class designator) or until no matching key can be found in the parsing table. The German words, now in rearranged word order, are looked up in the lexicon for English substitutions. Idioms that were unrecognized in the initial word arrangement but are reassembled in the parsing step (e.g., separable verbs) are replaced with the appropriate English equivalents, and alternate translations, based on neighboring semantic class designators or keywords present in a context register, are now substituted. The entire source document is thus translated from German to English one sentence at a time.

Performance Evaluation. Performance evaluation for this translation strategy includes assessing the efficiency of the parsing table and measuring the success of translation of potentially ambiguous terms. Since TRANSOFT is table-driven, its repertoire of idioms, alternate translations, and parsing formulas can be extended almost indefinitely, until an arbitrary level of translation quality is achieved. Carried to the extreme, each whole sentence could be translated as a separate "idiom" [2]. The obvious disadvantage in such a trivial strategy for automated translation is that the "idioms" and/or parsing formulas used in translating earlier parts of a document would almost never be reused for later parts of the document, so that creation of the lexicon and table would be virtually equivalent to manually translating the entire document. Thus we say that automated translator A is quantitatively more efficient than automated translator B if, to yield final translations of comparable quality, translator A *acquires* fewer table entries than translator B. An automated translator is said to acquire a new lexicon entry (or parsing formula) when it encounters the entry (or formula) for the first time. *Acquisition curves* for a translator and document can be obtained from the number of first occurrences of a lexicon entry (or parsing formula) as a function of distance through the document. For graphic representation, it is convenient to divide a document into deciles (10 percent intervals); for statistical analysis, the document can be divided in half.

A second measure of automated translator performance is the satisfactory translation of potentially ambiguous words in the source language that have two or more translations in the target language [28]. In a draft-quality transla-

tor, the available disambiguation mechanisms should resolve at least major differences in meaning. The TRANSOFT system employs three means for resolution of potentially ambiguous terms: (1) provision in the lexicon for multiple word idioms; (2) alternate translations of words depending upon the parts of speech (semantic class designators) of neighboring words in the source or rearranged text; and (3) alternate translations based upon a context register of subject matter keywords maintained by the program as the translation proceeds. Words that are likely to confuse TRANSOFT are those with multiple translations within the designated subject matter but without characteristic neighboring words. For example, the German word "Aufnahme" may be translated as "(ordinary or roentgenologic) photograph," "(clinic) registration," or "(biochemical) uptake," all in a medical context. In an effort to assess ambiguity resolution in our translation, all occurrences of 12 potentially ambiguous words were examined, and the frequency of resolvable and unresolvable ambiguities was determined.

RESULTS

The TRANSOFT medical document translator translated the entire computer-readable text of Adler's *Knochenkrankheiten* (*Bone Diseases*). Samples of the resulting translation, excerpted from the beginning, middle, and end of the document, are shown in **Table I.** The document contains 7,211 sentences, 118,604 words, 10,216 distinct words, and 844,715 characters. This represents an average of 16.4 words per sentence and 6.1 letters per word. The distribution of words by semantic class is shown in **Table II.** Nouns (N) were the most common semantic class, with 24,519 occurrences, followed by adjectives (A) with 16,059 occurrences. The least common semantic class was the one for foreign words or proper names (F), with 640 occurrences. The document required a total of 38,453 parsing formula look-ups of 2,791 distinct parsing formulas. This represents an average of 13.7 look-ups for each parsing formula used, 5.3 formulas used per sentence, and 0.4 new formulas acquired per sentence. The 20 most commonly used parsing formulas are shown in **Table III.** The entire book was translated in 16.9 hours during non-peak periods of computer activity (nights or weekends), an average of 7,025 words translated per hour.

The acquisition curves for the lexicon and parsing table are shown in **Figures 1** and **2.** For the lexicon, 25 percent of the words appear in the first 10 percent of the document, and 71 percent appear in the first half of the document ($p < 0.001$). A slight increase in number of new words near the end of the book corresponds to the introduction of additional terminology about bone tumors and cytophotometry. For the parsing table, 37 percent of the parsing formulas are used in the first 10 percent of the document, and 78 percent are used in the first half of the document ($p < 0.001$). The stereotypic nature of sentence construction is further illustrated by **Figure 3,** in which the acquisition curve for the first half of the document alone is compared with the acquisition curve for the second half of the document alone. These curves are very similar, suggesting a uniformity of grammatical style throughout the book.

Two classes of potential ambiguities were studied. The

TABLE II Distribution of Words by Semantic Class

Semantic Class	Number of Occurrences
. Period	7,211
, Comma	4,041
(Left parenthesis	3,150
) Right parenthesis	3,150
A Adjective or adverb	16,059
B Adverb only	4,756
C Conjunction	3,523
D Definite or indefinite article; demonstrative pronoun	15,174
F Foreign word or proper name	640
H Helping verb	5,622
I Interrogative or relative pronoun	1,484
N Noun	24,519
P Preposition	10,892
Q Pronoun	3,266
U Verb or participle	6,190
V Verb only	5,143
Z Number	3,784
Total	118,604

TABLE III 20 Most Commonly Used Parsing Formulas (E = noun phrase, R = prepositional phrase)

Rank	Key	Parsing Formula	Number of Occurrences
1	[DN	1[2ED3□N	1,478
2	DN	1ED2□N	1,304
3	PN	1RP2□N	1,228
4	[PN	1[2RP3□N	1,150
5	N(Z)	1N2□(3□Z4□)	1,147
6	DAN	1ED2□A3□N	1,133
7	PDN	1RP2□D3□N	720
8	AN	1[EA2□N	563
9	PDAN	1RP2□D3□A4N	561
10	IN	1RI2□N	520
11	[AN	1[2EA3□N	507
12	[EV[	1[2□E3V4[	437
13	BDN	1EB2□D3□N	383
14	PAN	1RP2□A3□N	356
15	[PNZ	1[2RP3□N4□Z	335
16	[RV[	1[2□R3V4[	333
17	DAAN	1D2□A3□A4N	328
18	AAN	1A2□A3N	298
19	BPN	1RB2□P3□N	268
20	[RH[	1[2□R3VH4[	253

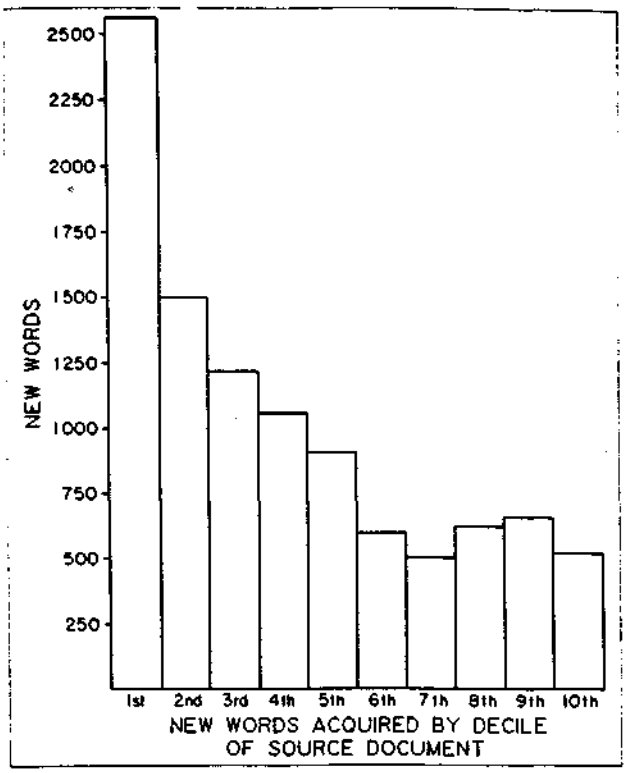


Figure 1. *Acquisition curve for the word lexicon. The document is divided into deciles along the abscissa, and the ordinate gives the number of first occurrences (acquisitions) that occur in each decile of the document. In a total of 10,216 word acquisitions, 2,557 (25 percent) new words appear in the first decile of the document, and 7,227 (71 percent) appear in the first half of the document. Significantly more acquisitions take place in the first half than in the last half of the document ($p < 0.001$). A slight increase in the number of new words near the end of the book corresponds to the introduction of additional terminology about bone tumors and cytophotometry.*

first, nouns at the beginning of German sentences, proved to be relatively minor. Five of 7,211 sentences began with a noun that had been placed in lower case by the preeditor, since the corresponding lower case word was another correctly spelled German word: three sentences beginning with "Teile" (parts) and two sentences beginning with "Schmerzen" (pain). These errors were easily corrected by a bilingual speaker. We then examined all occurrences of 12 potentially ambiguous words: six small words (als, am, das, der, die, sein) and six nouns (Aufnahme, Band, Mark, Praeparat, Seite, Zug). Results are shown in Table IV. All 1,761 occurrences of the potentially ambiguous small words could be resolved from a few surrounding words or semantic classes, but several of the 631 noun occurrences required human judgment to resolve the ambiguity. The single unresolvable occurrence of "Aufnahme" was this sentence (p 74):

"Die Aufnahme erfolgt hauptsaechlich durch das Trinkwasser."

"The uptake results primarily through the drinking water."

The translation of "Aufnahme" as "uptake" rather than "photograph" follows from the immediately preceding discussion of fluoride metabolism and the absurdity of making a photograph with drinking water. Likewise, a single occurrence of "Praeparat" required human judgment to distinguish between a gross or microscopic (slide) preparation. The word "Seite" occurred 335 times and required 16 human judgments, usually involving common sense for "eine Seite" ("one side" versus "one page"). In this book on bone diseases, "Band" (ligament or intervertebral disk, band) could always be translated correctly from the immediate context, and "Mark" (marrow) and "Zug" (traction) each had a single translation throughout the document, which was clear from the bone disease keywords present in the context register.

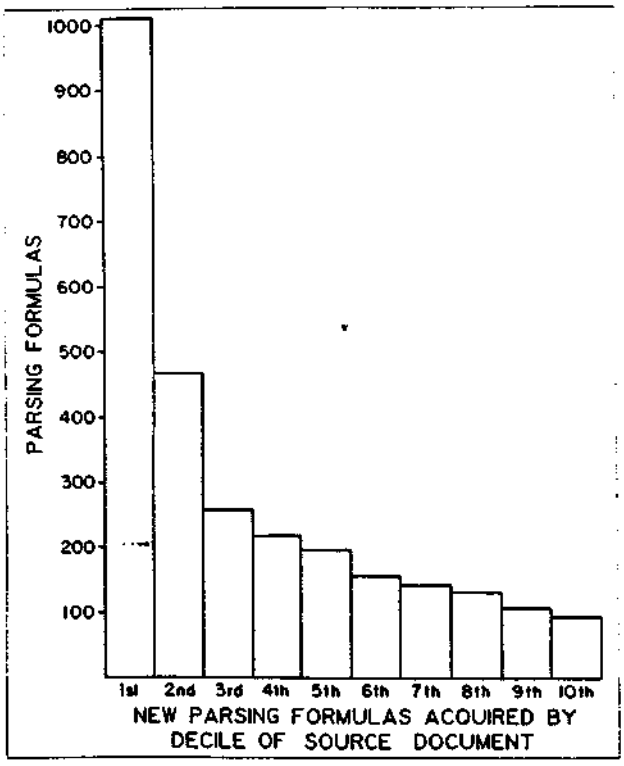


Figure 2. *Acquisition curve for the parsing table. The document is divided into deciles along the abscissa, and the ordinate gives the number of first occurrences (acquisitions) that occur in each decile of the document. In a total of 2,791 parsing formula acquisitions, 1,021 (37 percent) new parsing formulas appear in the first decile of the document, and 2,165 (78 percent) appear in the first half of the document. Significantly more acquisitions take place in the first half than in the last half of the document ($p < 0.001$).*

COMMENTS

Although draft-quality automated translators have recently become available commercially, it is not clear whether these translators will handle the specialized vocabulary and idioms of medical documents. Medical usage changes rapidly, and even recently published lexicons often lack the latest terminology. For this reason, an automated translator for medical text must allow the user to easily update both the vocabulary and the grammatical rules. TRANSOFT is such a table-driven medical document translator written in American National Standard MUMPS, for which the program source code has been published [17]. With TRANSOFT, the user has virtually complete control over the choice of semantic classes, default and alternate translations, idioms, and word rearrangement or parsing formulas. Additions and modifications to control tables can be made using either an interactive computer terminal or a communicating word processor, and the translator can be "fine-tuned" to produce a final translation of arbitrary quality. Fine tuning, however, is tedious and decreases the efficiency of the translator, since the translator is then more likely to need new idioms and parsing formulas that it has never encountered before as it reaches the later part of the document.

The TRANSOFT system and other practical automated translation systems currently in routine use employ the design principles of the Russian-to-English translation system developed at Georgetown University [13]. Other automated translators of similar design are in use at Oak Ridge National Laboratory and Wright-Patterson Air Force Base for translating Russian to English, at the Luxembourg headquarters of the European Economic Community for translating English to French, French to English, and English to Italian, and also at the Pan American Health

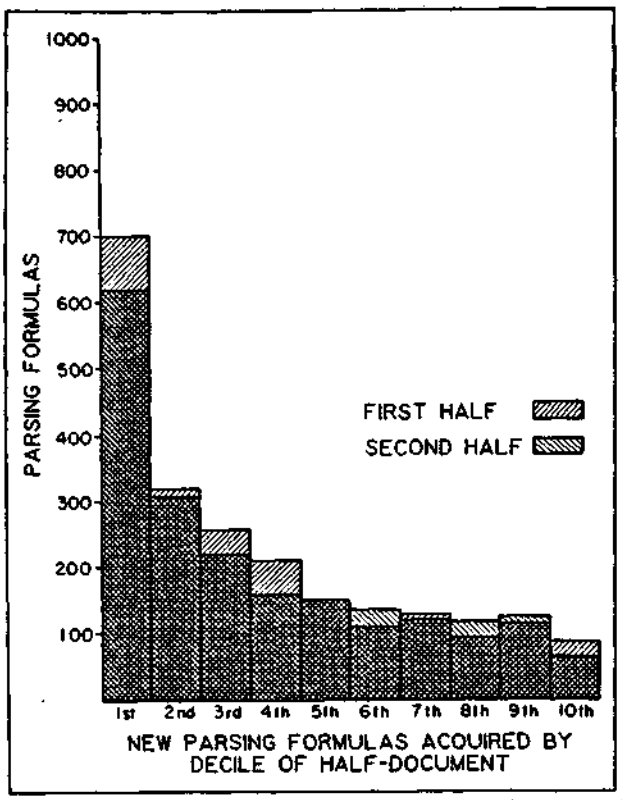


Figure 3. Acquisition curves for the parsing table applied to the first half of the document alone (▨) compared with the acquisition curve for the second half of the document alone (▩). Each half-document is divided into deciles along the abscissa, and the ordinate gives the number of first occurrences (acquisitions) that occur in each decile of the half-document. The stereotypic nature of sentence construction in the document as a whole is suggested by the similarity of these two document halves considered separately.

TABLE IV List of 12 Ambiguous Words

German Word	Total Occurrences	Default Translation	Alternate Translation	Number of Alternates	Number Requiring Judgment
als	339	as	than	69	0
am	143	to the	the most	71	0
, das	164	, which	, the	2	0
, der	131	, which	, the	3	0
, die	818	, which	, the	11	0
sein	166	to be	its	0	0
Aufnahme	35	photograph	uptake	3	1
Band/Baend	40	ligament	band	4	0
Mark	205	marrow	medulla	0	0
Praeparat	7	preparation	microscopic slide	2	1
Seite	335	page	side	43	16
Zug	9	traction	train	0	0
Total	2,392			206	18

Organization in Washington, for translating Spanish to English [5,13,29]. All these systems treat the text as a series of independent, unconnected sentences, and each sentence as a consecutive stream of words and idioms. They contain relatively little "understanding" such as is now being incorporated in the more recent prototype translation systems [6–9,12]. Despite these limitations, several Georgetown design systems have proved useful and cost-effective and constitute the majority of automated translation systems in routine use.

Although the task of entering new vocabulary and usage may seem unreasonably burdensome, the first-encounter rate for new table entries in the source document decreased rapidly. In the first 10th of the book, 25 percent of words and 37 percent of parsing formulas were encountered for the first time; however, in the last 10th of the book, only 5 percent of words and 3 percent of parsing formulas were encountered for the first time. This confirms our intuitive impression that medical text has a stereotyped vocabulary and usage that tends to repeat itself. The German vocabulary of 10,216 distinct words in this book of 118,604 words compares with an English vocabulary of 11,642 distinct words in a file of 7,000 consecutive autopsy reports with 923,657 words [26]. Thus, it seems likely that both the lexicon and the parsing tables generated for the initial document in a limited subject area will be largely applicable to all similar documents.

It remains controversial whether "fully automated high-quality translation" is a practical goal [28]. Our study does not resolve this question, since the output shown in Table I is clearly a draft-quality translation. At issue is whether a computer program, no matter how well-supplied with user tables, can satisfactorily resolve ambiguities [2]. We were surprised at how rarely ambiguities posed a serious problem within this specialized book about bone diseases. For example, "Mark" occurs 205 times and always means "marrow" (never "medulla" or "German mark"); "Zug" occurs nine times and always means "traction" (never "train" or "draft"). Likewise, "Aufnahme" occurs 35 times, with 32 default translations of "photograph" and three alternate translations of "uptake," only one of which could not be determined from a few neighboring words. These findings suggest that if the automated translator can determine the general subject matter from keywords in the context register, then serious ambiguities may be relatively rare.

The TRANSOFT system demonstrates that draft-quality, German-to-English medical translations can be obtained on a minicomputer from a computer-readable source document. After an initial acquisition process, the vocabulary and grammar as reflected in the lexicon and parsing tables settle into stereotyped patterns. Our findings do not support the idea that ambiguities pose a serious problem, as long as the appropriate contextual cues are utilized by the automated translator. As foreign language medical documents become increasingly available in computer-readable form through computerized typesetting, electronic publishing, and improved optical character recognition equipment, automated translation systems may provide a rapid and inexpensive means of obtaining draft translations [28,30,31].

DEDICATION

This paper is dedicated to the memory of the late Prof. Walter Sandritter, M.D. "Was ist das Schwerste von allem? Was dir das Leichteste dünket: Mit den Augen zu sehen, Was vor den Augen dir liegt." (What is the most difficult thing of all? That which would seem the easiest: to see with your eyes what actually lies before them) [32].

REFERENCES

1. Locke WN, Booth AD: Historical introduction. In: Locke WN, Booth AD, eds. Machine translation of languages. New York: John Wiley & Sons, 1955; 1–14.
2. Bar-Hillel B: Language and information. Selected essays on their theory and application. Reading, Massachusetts: Addison-Wesley, 1964; 153–218.
3. Automatic Language Processing Advisory Committee: Language and machines: computers in translation and linguistics (publication 1416). Washington: Division of Behavioral Sciences, National Academy of Sciences, National Research Council, 1966; 1–34.
4. Loh S-C: Machine translation: past, present, and future. ALLC Bulletin 1976; 4: 105–114.
5. Hutchins WJ: Progress in documentation. Machine translation and machine-aided translation. J Document 1978; 34: 119–159.
6. Wilks Y: An artificial intelligence approach to machine translation. In: Schank RC, Colby KM, eds. Computer models of thought and language. San Francisco: WH Freeman, 1973; 114–151.
7. Wilks Y: An intelligent analyzer and understander of English. Commun ACM 1975; 18: 264–274.
8. Wilks Y: A preferential, pattern-seeking, semantics for natural language inference. Artif Intell 1975; 6: 53–74.
9. Carbonell JG, Cullingford RE, Gershman AV: Steps toward knowledge-based machine translation. IEEE Trans PAMI 1981; 3: 376–392.
10. Wingert F: Lecture notes in medical informatics. In: Lindberg DAB, Reichertz PL, eds. Medical informatics. An introduction. New York: Springer-Verlag, 1981; 22–25.
11. Garfield E: Current comments. Artificial intelligence: using computers to think about thinking. Part 1. Representing knowledge. Current Contents 1983; 49: 5–17.
12. Schank RC, Childers PG: Experiments in artificial intelligence. Computerworld 1984; 10: 1–28.
13. Tucker AB Jr: A perspective on machine translation: theory

and practice. Commun ACM 1984; 27: 322–329.
14. Feigenbaum EA, McCorduck P: The fifth generation: artificial intelligence and Japan's computer challenge to the world. Reading, Massachusetts: Addison-Wesley, 1983.
15. Wa-bun ei-ya-ku system sho-hin ka. (Commercialized system for translation of Japanese sentences into English.) Asahi Shinbun, May 18, 1984.
16. Bulkeley WM: Computers gain as language translators even though perfect not they always. Wall Street Journal, February 6, 1985; 29.
17. Moore GW, Miller RE, Hutchins GM: Microcomputer translation for medical text: theorem verification for chapter two of Zeman's modal logic. Adv Math Comput Med (in press).
18. Adler C-P: Knochenkrankheiten. Diagnostik Makroskopischer, Histologischer und Radiologischer Strukturveränderungen des Skeletts. New York: Georg Thieme Verlag, 1983.
19. Bowie J, Barnett GO: MUMPS—an economical and efficient time-sharing system for information management. Comput Programs Biomed 1976; 6: 11–22.
20. Watanabe T, Ohsawa T, Suzuki T: A simple database language for personal computers. Commun ACM 1983; 26: 646–653.
21. Horowitz GL, Bleich HL: Paperchase: a computer program to search the medical literature. N Engl J Med 1981; 305: 924–930.
22. Robboy SJ, Altshuler BS, Chen HY: Retrieval in a computer-assisted pathology encoding and reporting system (CAPER). Am J Clin Pathol 1981; 75: 654–661.
23. Barnett GO, Justice NS, Somand ME, et al: COSTAR—a computer-based medical information system for ambulatory care. IEEE Proc 1979; 67: 1226–1237.
24. Barnett GO: The application of computer-based medical-record systems in ambulatory practice. N Engl J Med 1984; 310: 1643–1650.
25. Miller RE, Steinbach GL, Dayhoff RE: A hierarchical computer network: an alternative approach to clinical laboratory computerization in a large hospital. In: Proceedings of the Fourth Annual Symposium on Computer Applications in Medical Care, 1980; 505–513.
26. Moore GW, Hutchins GM, Miller RE: Strategies for searching medical natural language text: distribution of words in the anatomical diagnoses of 7000 autopsied patients. Am J Pathol 1984; 115: 36–41.
27. McGuire JF, Cooper RM: The Veterans Administration's approach to hospital automation. In: Proceedings of the Seventh Annual Symposium on Computer Applications in Medical Care, 1983; 76–79.
28. Winograd T: Computer software for working with language. Sci Am 1984; 251: 130–145.
29. Jordan SR, Brown AFR, Hutton FC: Computerized Russian translation at ORNL. J Am Soc Info Sci 1977; 28: 26–33.
30. Becker JD: Multilingual word processing. Sci Am 1984; 251: 96–107.
31. Walther von Alten J: Translators gain fluency. Multilingual word processors point to universal communication. InfoWorld 1984; 6: 35–37.
32. Sandritter W: Histopathologie. Lehrbuch und Atlas für Studierende und Ärzte, sixth ed. New York: FK Schattauer Verlag, 1975.

Reprinted from the July issue of **The American Journal of Medicine**, A Yorke Medical Journal, Published by Technical Publishing a Division of Dun-Donnelley Publishing Corporation, a Company of the Dun & Bradstreet Corporation, 875 Third Avenue, New York, N.Y., 10022.