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 1980s 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 theoric 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
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 TRANSFO, 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 TRANSFO 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 TRANSFO 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 Krankenheiten 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 Aa, Oe, Ue, aa, 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 nonhyphenated 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 TRANSFO system. Thus the first two sentences of Chapter One:

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

Lexicon. A lexicon of words and idioms is one of two external tables of language-specific control information used by the TRANSFO 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 Krankenheiten source file by collating all character strings bounded on either side by a blank, a task easily performed in MUMPS using the SPECTE 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), entzg (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)
E = 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 = proposition, 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 TRANSFORT system. Parsing formulas are applied recursively by TRANSFORT 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 Körper lässt sich durch einen in der Mitte geführten Schritt zentrieren.

The body allows itself to pass through a section passed in the middle.

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

\[ \begin{align*}
 & \text{DPNUN} \\
 & \text{DNN} \\
 & \text{(old parsandum substring)} \\
 & \text{DN} \\
 & \text{(new parsandum substring)} \\
\end{align*} \]

or more compactly by the formula \( 1D4D64P50D60N9U2N \) \( \frac{1Q2D3P3D4D5S}{} \) \( \frac{1[2D3O4MVH5Q5QV7]}{} \), 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:

\[ \text{einen Schritt} \]

\[ \text{DND} \]

\[ \text{a section} \]

\[ \text{(exposed)} \]

\[ \text{in der Mitte} \]

\[ \text{UPD} \]

\[ \text{passed in the middle} \]

where "einen Schritt 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 TRANSFORT 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:

\begin{align*}
\text{Parsandum} & = [DNHQDPDUNUV] \\
\text{Parsing Formula} & = 1D4D64P50D60N9U2N \\
& = 1Q2D3P3D4D5S \\
& = 1[2D3O4MVH5Q5QV7] \\
& \text{(complete)}
\end{align*}

The final, English word order is then determined as:

\[ \text{Der Körper lässt sich durch einen in der Mitte gehaltenen Schritt zentrieren.} \]

\[ \text{The body allows itself to pass through a section passed in the middle.} \]

The parsing table for this German-to-English translation was generated incrementally by having TRANSFORT 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 TRANSFORT 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 Hin-
sicht eine zentrale Stellung ein. Es verleiht jedem Lebewe-
gen seine spezifische Körperform und ist bestimmend
fur die Architektontik des Körpers. Gleichzeitig wird die
Grosse des Individuums entscheidend durch das Skelett
gepaegt. Diese Formgebung durch das Skelett ist propor-
tional und symmetrisch angelegt, wobei die Groesse der
einzelnen Skeletzteile dem Gesamtskelett angeglichen ist.

Osteochondrom. Osteochondroms sind die weitaus häufig-
gsten gutartigen Knochenwucherungen. Unter den benignen
Knochentumoren haben sie einen Anteil von 40 %. Es han-
delt sich um eine knochenerne Neubildung, die von einer
breiten Kappe aus hyalinem Knorpelgewebe u eberzogen
ist und sich von der Knochenbasisfläche pittoformig in die
umgebende Weichteilhülle vorwölbt. Da diese Tumoren nur
langsamer an Groesse zunehmen, machen sie oft erst spät
durch eine Schwellung auf sich auffallen.

Traumatische Meniskusläsion. Hierbei handelt es sich um
eine mechanische Verrenkung eines gesunden Meniskus,
meist infolge eines sportlichen oder beruflichen Unfalls.
Von dieser häufigen Verletzung ist vor allem der mediale
Meniskus betroffen (hochmässig häufiger als der laterale
Meniskus). Man unterscheidet einen vollständigen oder
teilweisen Meniskusrisss von einem häufigeren Substan-
riss, bei dem der Meniskus gespalten wird. Die Verletzung
erfolgt bei plötzlicher Streckung und gleichzeitiger Rotati-
on im Kniegelenk.

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 architecture of the
body. Simultaneous the size of the individual becomes
pressed upon decisive through the skeleton. This giving of
form through the skeleton is laid out proportional and symmet-
ric, whereby is adjusted the size the individual skeletal parts
to the skeleton as a whole.

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 hya-
in 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.

Traumatic meniscus lesion. Hereat one is dealing with a me-
chanical tearing of a healthy meniscus most as a result of a
sporadly 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 distin-
guishes a complete or partial meniscal tear from a more fre-
frequent substance tear, at which is cleft the meniscus. The inju-
ry results at sudden stretching and simultaneous rotation in
the knee joint.

The parsing table and measuring the success of translation of
potentially ambiguous terms. Since TRANSFOLD 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 "idi-
oms" 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, transla-
tor 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 occu-
crences of a lexicon entry (or parsing formula) as a function
of distance through the document. For graphic representa-
tion, it is convenient to divide a document into decades (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 transla-
tions 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,” “(clinical) 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 Knochenerkrankheiten (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 the 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

<p>| TABLE II Distribution of Words by Semantic Class |</p>
<table>
<thead>
<tr>
<th>Semantic Class</th>
<th>Number of Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>7,211</td>
</tr>
<tr>
<td>Comma</td>
<td>4,041</td>
</tr>
<tr>
<td>(Left parenthesis</td>
<td>3,150</td>
</tr>
<tr>
<td>) Right parenthesis</td>
<td>3,150</td>
</tr>
<tr>
<td>A Adjective or adverb</td>
<td>16,059</td>
</tr>
<tr>
<td>B Adverb only</td>
<td>4,758</td>
</tr>
<tr>
<td>C Conjunction</td>
<td>3,523</td>
</tr>
<tr>
<td>D Definite or indefinite article:</td>
<td>15,174</td>
</tr>
<tr>
<td>demonstrative pronoun</td>
<td></td>
</tr>
<tr>
<td>F Foreign word or proper name</td>
<td>640</td>
</tr>
<tr>
<td>H Helping verb</td>
<td>5,622</td>
</tr>
<tr>
<td>I Interrogative or relative pronoun</td>
<td>1,454</td>
</tr>
<tr>
<td>N Noun</td>
<td>24,519</td>
</tr>
<tr>
<td>P Preposition</td>
<td>10,892</td>
</tr>
<tr>
<td>Q Pronoun</td>
<td>3,266</td>
</tr>
<tr>
<td>U Verb or participle</td>
<td>6,190</td>
</tr>
<tr>
<td>V Verb only</td>
<td>5,143</td>
</tr>
<tr>
<td>Z Number</td>
<td>3,764</td>
</tr>
<tr>
<td>Total</td>
<td>118,604</td>
</tr>
</tbody>
</table>

<p>| TABLE III 20 Most Commonly Used Parsing Formulas (E = noun phrase, R = prepositional phrase) |</p>
<table>
<thead>
<tr>
<th>Rank</th>
<th>Key</th>
<th>Parsing Formula</th>
<th>Number of Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[DN]</td>
<td>1[2EDO2N]</td>
<td>1,478</td>
</tr>
<tr>
<td>2</td>
<td>DTN</td>
<td>1EDO2DN</td>
<td>1,304</td>
</tr>
<tr>
<td>3</td>
<td>PN</td>
<td>1RPZ2DN</td>
<td>1,228</td>
</tr>
<tr>
<td>4</td>
<td>PN</td>
<td>1[2RP3DN]</td>
<td>1,150</td>
</tr>
<tr>
<td>5</td>
<td>N(2)</td>
<td>1N2O3DZ24D</td>
<td>1,147</td>
</tr>
<tr>
<td>6</td>
<td>DUN</td>
<td>1EDO2A3DN</td>
<td>1,133</td>
</tr>
<tr>
<td>7</td>
<td>PDN</td>
<td>1RPZ2D3DN</td>
<td>720</td>
</tr>
<tr>
<td>8</td>
<td>AN</td>
<td>1EA2DN</td>
<td>563</td>
</tr>
<tr>
<td>9</td>
<td>PDAn</td>
<td>1RP22D3O2A4N</td>
<td>561</td>
</tr>
<tr>
<td>10</td>
<td>IN</td>
<td>1RIZ2DN</td>
<td>520</td>
</tr>
<tr>
<td>11</td>
<td>AN</td>
<td>1EA23DN</td>
<td>507</td>
</tr>
<tr>
<td>13</td>
<td>BDN</td>
<td>1EB2D3DN</td>
<td>383</td>
</tr>
<tr>
<td>14</td>
<td>PAN</td>
<td>1RP2A3DN</td>
<td>356</td>
</tr>
<tr>
<td>16</td>
<td>RV</td>
<td>1[2RZ3Y4]</td>
<td>333</td>
</tr>
<tr>
<td>17</td>
<td>DAA2N</td>
<td>1D2O3A3DN</td>
<td>328</td>
</tr>
<tr>
<td>18</td>
<td>AAN</td>
<td>1A2D32N</td>
<td>298</td>
</tr>
<tr>
<td>19</td>
<td>BN</td>
<td>1RB2D3PN</td>
<td>288</td>
</tr>
<tr>
<td>20</td>
<td>[RH]</td>
<td>1[2OR3ZH]</td>
<td>253</td>
</tr>
</tbody>
</table>
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 "Telle" (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 hauptsächlich 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 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.](image1)

![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 1,761 (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).](image2)
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 translation 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 reordering 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 employment may be designed to use 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 (a) compared with the acquisition curve for the second half of the document alone (b). 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.](image)

**TABLE IV** List of 12 Ambiguous Words

<table>
<thead>
<tr>
<th>German Word</th>
<th>Total Occurrences</th>
<th>Default Translation</th>
<th>Alternate Translation</th>
<th>Number of Alternates</th>
<th>Number Requiring Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>als</td>
<td>336</td>
<td>as</td>
<td>than</td>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td>am</td>
<td>143</td>
<td>to the</td>
<td>the most</td>
<td>71</td>
<td>0</td>
</tr>
<tr>
<td>das</td>
<td>164</td>
<td>which</td>
<td>the</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>der</td>
<td>131</td>
<td>which</td>
<td>the</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>die</td>
<td>818</td>
<td>which</td>
<td>the</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>sein</td>
<td>168</td>
<td>to be</td>
<td>its</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Aufnahme</td>
<td>35</td>
<td>photograph</td>
<td>uptake</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Band/Baend</td>
<td>40</td>
<td>ligament</td>
<td>band</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Mark</td>
<td>205</td>
<td>marrow</td>
<td>medulla</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Praeparat</td>
<td>7</td>
<td>preparation</td>
<td>microscopic</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Seite</td>
<td>335</td>
<td>page</td>
<td>side</td>
<td>43</td>
<td>16</td>
</tr>
<tr>
<td>Zug</td>
<td>9</td>
<td>traction</td>
<td>train</td>
<td>43</td>
<td>16</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,392</strong></td>
<td></td>
<td></td>
<td><strong>208</strong></td>
<td><strong>18</strong></td>
</tr>
</tbody>
</table>
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 novel 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 TRANSTOF 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 dunkt: 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


