Methods of Information in Medicine


**G. W. Moore, U. N. Riede, R. A. Polacsek, R. E. Miller, G. M. Hutchins**

# Group Theory Approach to Computer Translation of Medical German*

*(From the Departments of Pathology and Laboratory Medicine, the Welch Medical Library of The Johns Hopkins Medical Institutions, Baltimore, Maryland, and the Department of Pathology of Freiburg University School of Medicine, Freiburg i. Br., F.R.G.)*

**Summary**

Computer translators have been studied for almost four decades, but recent advances in speed and storage capabilities have made such translators accessible to small computer users. We obtained the computer typesetting file for a German language medical textbook and wrote computer software sufficient to obtain a draft quality English language translation of the entire book, at a speed of 9,671 words per hour. This translator uses two external tables, namely a word and idiom list and a list of grammatical rules, which completely specify the behavior of the translator. The grammatical rule table satisfies the properties of a mathematical group, and the inverse operation for this group allows one in principle to convert this German to English translator into an English to German translator. For the larger problem of creating multilingual computer translators, the group theory inversion property may allow one to substantially reduce the effort of creating a separate translator for each language pair.
Future development of computer translators will depend upon the wider availability of computer-readable documents and will be aided by use of vocabulary and grammatical rule tables with group theory properties which permit the invertability between language pairs.

*Key-Words:* Group Theory, Computer Translation, German Natural Language, Medical Translation

**Die Anwendung des gruppentheoretischen Prinzips für die Computerübersetzung deutschen medizinischen Schrifttums ins Englische**

Beinahe vier Jahrzehnte lang sind Computerprogramme zum Zweck der Sprachübersetzung entwickelt worden, aber erst das raschere Arbeitstempo und die größere Speicherkapazität der neuen Geräte haben solche Übersetzerprogramme den Benutzern von Mini- und Mikrocomputern zugänglich gemacht. Wir erhielten den auf Magnetband gespeicherten, für die Setzmaschine bestimmten Text eines deutschsprachigen medizinischen Lehrbuchs und entwikkelten ein Computerprogramm, welches das ganze Buch mit einer Geschwindigkeit von 9671 Wörtern pro Stunde als ersten Entwurf übersetzte. Dieses Programm benutzt zwei externe Tabellen, nämlich eine Wörterliste einschließlich idiomatischer Terms und Fachausdrücke und eine Tabelle grammatikalischer Regeln. Diese beiden Tabellen allein bestimmen die Arbeitsweise des Programms. Die grammatikalische Regeltabelle besitzt die Eigenschaften einer mathematischen Gruppe, deren inverse Operation es prinzipiell ermöglicht, dieses deutsch-englische Übersetzerprogramm in einen englisch-deutschen Übersetzer zu verwandeln. Ein wichtiges Ziel wäre es, einen mehrsprachigen Computerübersetzer zu entwickeln. Die Inversion, eine Eigenschaft der Gruppentheorie, würde es ermöglichen, den Zeitaufwand für die Entwicklung mehrsprachiger Übersetzer erheblich herabzusetzen.
Die zukünftige Weiterentwicklung von Computerübersetzern wird von der erhöhten Verfügbarkeit computerlesbarer Dokumente abhängen, wird aber auch durch den Gebrauch von Vokabellisten und Sprachregeltabellen unterstützt, die aufgrund ihrer gruppentheoretischen Eigenschaften die Inversion zwischen Sprachpaaren erlauben.

*Schlüssel-Wörter:* Gruppentheorie, Computerübersetzung, deutsche Umgangssprache, Übersetzung medizinischer Texte

## Introduction

Computer translators have been studied for almost four decades [4, 8, 12, 14–16, 24, 25], but recent advances in speed and storage capabilities have made such translators available for practical translation problems on small computers [2, 6, 7, 10, 13, 21, 22, 26, 27, 28]. Our laboratory has introduced a German to English medical document translator which can be implemented on a minicomputer or even an appropriately configured microcomputer [18–20]. TRANSOFT is a table-driven German to English medical document translation system written in the American National Standard MUMPS programming language which was used to generate a draft quality English translation of a German language textbook [1]. The behavior of this translator is completely specified by the vocabulary and grammatical rule lists.

Several strategies have emerged in efforts to generalize a given computer translator (here, German to English) to other language pairs. The most straightforward strategy is to regard each language pair as a separate translation problem, with no method for generalization to other language pairs. Using this strategy, multilingual translation among the eight official languages of the European Economic Community (English, German, French, Italian, Dutch, Danish, Spanish, Greek) would require $8 \times 7 = 56$ separate translators [22]. A second strategy involves the construction of a language-free representation, or interlingua, as a general medium around which analysis of the source language and synthesis of the target language could be focused. It has been proposed that such an interlingua might be formalized as Chomsky transformational grammars or as augmented transitional networks. A third strategy

* Supported by Grant LMO3651 from the National Library of Medicine

has been suggested in which the computer would attempt to understand what it translates, using semantic information stored in a universal knowledge base [22]. In the present report, we suggest that the effort of creating a separate translator for each language pair can be significantly reduced, using the operations of mathematical group theory to invert a given translator.

A group, ( $\mathfrak{G}$ , $\circ$), consists of a set of objects, $\mathfrak{G}$ , and an operation, $\circ$ [3]. Each pair of objects in $\mathfrak{G}$ may be combined using operation $\circ$, and the result must be an object in $\mathfrak{G}$ . A group must also satisfy the property of associativity, i.e., $(a \circ b) \circ c = a \circ (b \circ c)$; there must be a special object in $\mathfrak{G}$ , the identity, z, with the property that $a \circ z = z \circ a = a$; and each a in $\mathfrak{G}$ must have an inverse, denoted $a^{-1}$, such that $a \circ a^{-1} = a^{-1} \circ a = z$. The group of integer addition, with operation +, identity z = 0, and inverse consisting of negation, is a familiar group. In this report we demonstrate that the set of grammatical rules for the TRANSOFT German to English medical document translator is a mathematical group. The demonstration of an inverse operation suggests that any TRANSOFT grammatical rule table can be inverted to obtain a grammatical rule table for the reverse language pair.

## Methods

A recent German language medical text was made available to us in computer-readable form by the publisher, Georg Thieme Verlag [1]. The full manuscript, excluding footnotes, table headings, and figure legends, was written from a 9-track magnetic tape to an American Standard Code for Information Interchange (ASCII) text file 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 preediting and then translated in its entirety from German into 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 (*M*assachusetts General Hospital *U*tility *Multi*Programming *S*ystem), which is widely used for medical information processing, was chosen for TRANSOFT programs because of its powerful character string operators and its string-subscripted arrays with implicit-sorting [5, 11, 17, 23].

*Computer Preediting.* Preliminary computer editing and reformatting of the raw text file was carried out to provide a standardized document for the subsequent translation steps, using rules which 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, ue and ss, respectively. Punctuation was reduced to commas, periods, semicolons, colons, exclamation points, question marks, single and double quotation marks, square and curly brackets, and parentheses. Each phrase delimited by dashes—such as this one—was enclosed in parentheses. Each hyphenated word was converted to a single nonhyphenated word. Terms containing numeric characters were marked with a "Chinese period" ($\circ$), and decimal numbers were expressed in American format using the Chinese period, e.g. 27$\circ$4 rather than 27,4. Punctuation characters other than 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 a lower case to simplify the later processing of nouns (which begin with an upper case character in German). Using 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 preedited text file then served as the source document for all subsequent processing by the TRANSOFT system.

*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 the present translation, an initial German word list was generated from the source document by collating all character strings bounded on either side by a blank. This list was then expanded to include additional noun and adjective declensions and verb conjugations, including separable verb forms. Potential idioms were obtained from a list of all word pairs or word triples occurring in the source document, listed in descending order of frequency. Words were accepted as final lexicon entries by a bilingual speaker who assigned a default translation and a syntactic-semantic class designator [9] to each entry, using the 21 classes listed below. Each word in the lexicon was also assigned any number of alternate translations which are dependent upon the classes of neighboring words. Most of the syntactic-semantic classes represent punctuation or ordinary parts of speech, although some reflect the unique requirements of computer translation. For example, U is an ambiguous part of speech commonly encountered in German, and Z represents words often encountered in scientific documents which require special processing. Each sentence begins with [ and terminates with ] (replacement for period). The concealment box, □, is a special character required by the TRANSOFT parsing formulas (see below).

, – comma
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).
E – noun phrase (= DN).
H – helping verb, e.g., sein (be), haben (have), werden (become).
I – interrogative or relative pronoun, e.g., welcher (which), warum (why).
J – preposition and determiner (= PD).
N – noun, e.g., Anwendung (application), Auftreten (appearance).
P – preposition, e.g., auf (upon), bei (at).
R – prepositional phrase (= PN)
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).
W – subject and verb (= QN).
Y – negation, e.g., nicht (not), kein (none).
Z – number or formula, e.g., eins (one), zwei (two), etc., word containing a numeric character, measurement, mathematical symbol, or foreign word; misspellings are assigned by default to class Z.
[ – left bracket (start character)
] – right bracket (stop character)
□ – concealment box

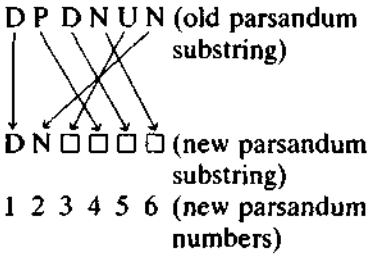
*Parsing Tables.* A parsing table of word rearrangement instructions, or parsing formulas, is the second translation table used by the TRANSOFT system. Parsing formulas are applied recursively by TRANSOFT to transform a sentence in German (source) word order to its corresponding English (target) word 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 computer translation systems [7, 21, 22, 24, 25]. In the parlance of transformational grammars, the lexicon entries are "terminals", the syntactic-semantic class designators are "nonterminals", the "initial sentence" is "[]", the parsing table is a "grammar", and the TRANSOFT parsing algorithm belongs to the class of "reduction grammars" [8, 10, 26]. An unparsed or incompletely parsed sentence in the source language can be represented by the consecutive sequence of syntactic-semantic class designators, called a parsandum, for that sentence. A parsing formula consists of a key (i.e., sequence of 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 clause has a parsandum of [DPDNUNV]:

, der das in die Gefaesse applizierte Kontrastmittel aufweist.

[ D P D N U N V ]

, which the into the vessels applied contrast medium exhibits.

Note that the syntactic-semantic class designator for the idiom ", der" is "[" and the class designator for "." is "]". 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:




This arrow diagram tells the sentence parser how to rearrange the substring from German to English word order and which part of the substring, now in English word order, should be concealed from the parser to prevent confusion in subsequent steps. In this example,

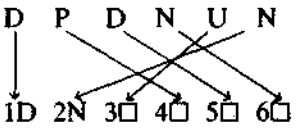
das Kontrastmittel
D N
the contrast medium

is retained for subsequent parsing steps. On the other hand, this substring:

applizierte in die Gefaesse
□ □ □ □
applied into the vessels

which is now in English word order, is concealed behind Kontrastmittel (contrast medium) in subsequent parsing steps by use of the concealment box designator.

Since currently available word processors do not readily handle arrow diagrams, this diagram can be transformed into an unambiguous notation which rests on a single line. The new parsandum numbers are pulled up to the left of the new parsandum designators:




The arrows are then pulled up to the left of the old parsandum designators:

1D←D4□←P5□←D6□←N3□←U
2N←N

Arrows and blanks are removed, and repeated letters are replaced by a single letter:

1D4□P5□D6□N3□U2N

The arrow diagram can readily be reconstituted from this unambiguous single line notation.

During iterative processing of each sentence by the TRANSOFT program, either the entire parsandum or else 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 at least one concealment box [18]. This mathematical property is akin to the "non-shortening" property of generative transformational grammars [10]. For the present example, the final English word order is obtained in three parsing steps:

| | *Parsandum* |
|---|---|
| 1. | [ D P D N U N V ] |
| 2. | [ D N V ] |
| 3. | [ V ] |

*Parsing Formula*

1. 1D4 □ P5 □ D6 □ N3 □ U2N
2. 1[3 □ D4 □ N2V5]
3. (complete)

, der aufweist das Kontrastmittel applizierte in die Gefaesse.

[ V D N U P D N ]

, which exhibits the contrast medium applied into the vessels.

The parsing table for the present German to English translator 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 for the entire source document. A sample translation obtained by TRANSOFT is shown in Table 1.

## Group Theory

In mathematics, a group ( $\mathfrak{G}$ , $\circ$), consists of set of objects, $\mathfrak{G}$ , and an operation, $\circ$ [3]. Any pair of objects in $\mathfrak{G}$ may be combined using operation $\circ$, and the result must be an object in $\mathfrak{G}$ (closure property). The set $\mathfrak{G}$ might consist of integers, real numbers, or other abstract objects. Two familiar group operations are addition (operation +) and multiplication (operation ×). A group must also satisfy the property of associativity, i.e., $(a\circ b)\circ c = a\circ(b\circ c)$; there must be a special object in $\mathfrak{G}$ , the identity, z, with the property that $a\circ z = z\circ a = a$; and each a in $\mathfrak{G}$ must have an inverse, denoted $a^{-1}$, such that $a\circ a^{-1} = a^{-1}\circ a = z$. The group of integer addition, with $\mathfrak{G}$ the set of all integers, operation +, identity $z = 0$, and inverse consisting of negation, is a familiar group. That is, if a, b, and c are integers, then a+b is also an integer (closure); (a+b)+c = a+(b+c) (associativity); a + 0 = 0 + a = a (identity); and a+(−a) = (−a)+ a = z (inverse). In the present report, the set $\mathfrak{G}$ is the set of parsing formulas, each expressed as an "offset vector", or simply "offset". An offset specifies the number of places rightward or leftward to move each sentence element in order to obtain the rearranged sentence. The operation, $\circ$, corresponds to the successive application of two offsets to a given sentence. Thus, if A is a sentence and $p,q \in \mathfrak{G}$ are offsets, then p(A) represents offset p applied to sentence A and $(q\circ p)(A)$ represents offset p, and then offset q, applied to sentence A. The most important result of a group theory approach to the TRANSOFT computer translator is the existence of an inverse property, which suggests that many of the parsing formulas needed to build an English to German translator are already implicit as inverses of German to English parsing formulas.

Since TRANSOFT is a sentence-by-sentence translator, we focus our attention on a single sentence (or clause) containing n members. It is convenient to number the sentence elements (words or idioms) consecutively from 1 to n. In the present example:

, der das in die Gefaesse applizierte Kontrastmittel aufweist.

1 2 3 4 5 6 7 8 9
[ D P D N U N V ]

, which the into the vessels applied contrast medium exhibits.

The sentence (initial parsandum) is designated by an upper case letter at the beginning of the alphabet, such as $A = (A_1,A_2,\ldots,A_9)=(1,2,\ldots,9)$. Each offset which may be applied to A (or other parsanda) is represented by a lower case letter in the middle of the alphabet, such as p. An offset is an n-vector which may contain integers between −n and +n. The kth element in the offset states how many places rightward (positive value) or leftward (negative value) to move each element in the sentence, A. If $p_k=0$, then the kth element in A does not move. Thus p = (0, 0,+2,+2,+2,−2,−4, 0, 0) applied to A specifies that elements 1, 2, 8, and 9 should remain unchanged; elements 3, 4, and 5 should each be moved rightward two places; element 6 should be moved leftward 2 places and element 7 leftward by 4 places. That is:

A = (1 2 3 4 5 6 7 8 9)
p = (0 0 +2 +2 +2 −2 −4 0 0)
p(A) = (1 2 7 6 3 4 5 8 9)

*Table 1* Sample German to English translation (ref. [1], p. 36)

| | |
|---|---|
| Osteogenesis imperfecta. Es handelt sich um einen allgemeinen Defekt des Mesenchyms, der zu einer fehlerhaften Kollagensynthese und einer unzulänglichen Knochenbildung führt. Die Osteogenesis imperfecta ist eine erbliche Skeletterkrankung, die durch eine starke Knochenbrüchigkeit gekennzeichnet ist. Ursache ist ein allgemeiner Defekt des Mesenchyms, der zu einer fehlerhaften Kollagensynthese und einer unzulänglichen Knochenbildung führt, so daß außer der Knochenbrüchigkeit eine Hypermobilität der Gelenke, hellblaue Skleren, eine Otosklerose, Zahnanomalien und eine pergamentartige Verdünnung der Haut vorkommen. Die Osteogenesis imperfecta congenita Vrolik führt bereits intrauterin oder bald nach der Geburt zum Tod. Bei der Osteogenesis imperfecta tarda Lobstein treten Frakturen erst im späteren Kindesalter und Jugendlichenalter auf. Die Patienten haben eine normale Lebenserwartung. | Osteogenesis imperfecta. One is dealing with a general defect of the of mesenchyme, which leads to a mistaken collagen synthesis and an insufficient bone formation. The osteogenesis imperfecta is a hereditary skeletal disease, which is characterized by a strong bone fracturability. Cause is a general defect of the of mesenchyme, which leads to a mistaken collagen synthesis and an insufficient bone formation, so that occur besides the bone fracturability a hypermobility the joints, bright blue scleras, a otosclerosis, tooth anomalies and a parchment-like thinning the skin. The congenital osteogenesis imperfecta of Vrolik leads already intrauterine or soon after the birth to the death. At the osteogenesis imperfecta tarda of Lobstein appear fractures first in the later pediatric age group and adolescent age. The patient have a normal life expectancy. |

Note that the sum of elements in p add up to 0; and p(A), like A, contains all the integers between 1 and n, albeit in different order. Sample calculations involving offsets are shown in Table 2. The set of offsets constitutes a non-Abelian (non-commutative) group. The identity is the zero offset, $z = (0,\ldots,0)$; and every offset has an inverse.

For the following definitions and theorems, let [1, n] denote the closed interval between 1 and n, inclusive, which contains all and only integers. Let $\mathfrak{A}$ be the set of all ordered arrangements of [1,n]. Let $N = [-n,n]$.

*Definition 1 (Offset):* Let $p \in N \times \ldots \times N$ (n times), and $p = (p_1,\ldots,p_n)$. Then p is an offset, i.e., $p \in \mathfrak{S}$, if and only if for every $j \in [1,n]$ there exists a unique $k \in [1,n]$ such that $k+p_k = j$. For $A \in \mathfrak{A}$, $(p(A))_{k+pk} = A_k$.

The sum of elements in every offset is 0.

*Theorem 1:* Let p be an offset.

Then $\sum_{k=1}^{n} p_k = 0$.

*Proof.* By Definition 1, for every $j \in [1,n]$ there exists a unique $k \in [1,n]$ such that $k + p_k = j$. Therefore k assumes each value between 1 and n, so that

$$\sum_{k=1}^{n} (k+p_k) = \sum_{j=1}^{n} j \text{ and}$$

$$\sum_{k=1}^{n} p_k = \sum_{j=1}^{n} j - \sum_{k=1}^{n} k = 0.$$

If A is an ordered arrangement and p is an offset, then p(A) is also an ordered arrangement.

*Theorem 2:* Let p be an offset and $A \in \mathfrak{A}$. Then there exists a $B \in \mathfrak{A}$ such that $B = p(A)$.

*Proof.* Let $B = p(A)$ and consider any $j \in [1,n]$. Then by Definition 1 there exists a unique $k \in [1, n]$ such that $k+p_k = j$, and $B_j = B_{k+pk} = [p(A)]_{k+pk} = A_k \in [1, n]$.

The inverse of offset p is obtained by setting the jth element in the inverse equal to negative the kth element in p, where $k+p_k = j$.

*Table 2* Sample calculations involving sentences and offsets [Sentences A, p(A), z(A), (p∘q)(A), and (q∘p)(A), are all members of $\mathfrak{A}$. Offsets p, $p^{-1}$, z, q, p∘q, and q∘p are all members of set $\mathfrak{S}$. Note that (q∘p) ≠ (p∘q), demonstrating that group ($\mathfrak{S}$, ∘) is non-Abelian (non-commutative)].

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | [ | D | P | D | N | U | N | V | ] | (German word order) |
| A = | (1, | 2, | 3, | 4, | 5, | 6, | 7, | 8 | 9) | |
| p = | (0, | 0, | +2, | +2, | +2, | −2, | −4, | 0 | 0) | |
| p(A) = | (1, | 2, | 7, | 6, | 3, | 4, | 5, | 8, | 9) | |
| $p^{-1}$ = | (0, | 0, | +4, | +2, | −2, | −2, | −2, | 0, | 0) | |
| $z=p^{-1}\circ p$ = | (0, | 0, | 0, | 0, | 0, | 0, | 0, | 0, | 0) | |
| z(A) = | (1, | 2, | 3, | 4, | 5, | 6, | 7, | 8, | 9) | |
| q = | (0, | +1, | +1, | +1, | +1, | +1, | +1, | −6, | 0) | |
| (p∘q) = | (0, | +3, | +3, | +3, | −1, | −3, | +1, | −6, | 0) | |
| (p∘q)(A) = | (1, | 8, | 6, | 5, | 2, | 3, | 4, | 7, | 9) | |
| q∘p = | (0, | +1, | +3, | +3, | +3, | −1, | −3, | −6, | 0) | |
| (q∘p)(A) = | (1, | 8, | 2, | 7, | 6, | 3, | 4, | 5, | 9) | |
| | [ | V | D | N | U | P | D | N | ] | (English word order) |

*Definition 2 (Inverse):* Let p be an offset. Then q is the inverse of p if and only if for every $j \in [1, n]$ and $k+p_k = j$, $q_j = -p_k$.

The inverse of an offset is itself an offset, and the inverse of the inverse of an offset is itself.

*Theorem 3:* Let q be the inverse of p. Then (*i*) q is an offset, and (ii) p is the inverse of q.

*Proof. Part (i).* Consider any $k \in [1,n]$; by Definition 1, there exists a unique $j \in [1,n]$ such that $j+p_j = k$. By Definition 2, $q_k = -p_j$. Adding k to both sides of the equation and substituting gives: $k+q_k = k-p_j = (j+p_j)-p_j = j$. Since $j,k \in [1,n]$ are in one to one correspondence, for each j there exists a k. By Definition 1, q is an offset.

*Part (ii).* By part (i), q is an offset. Consider any $k \in [1, n]$. By Definition 1, there exists a unique $j \in [1,n]$ such that $j+p_j = k$. By Definition 2, $q_k = -p_j$ and $j = k-p_j = k+q_k$. Since $j,k \in [1,n]$ are in one to one correspondence, for each j there exists a k. By Definition 2, p is the inverse of q.

The composition of two offsets, denoted $r = q \circ p$, is obtained by the expressions $i = j+q_j$, $j = k+p_k$, and $r_k = q_j+p_k$.

*Definition 3 (Composition):* Let q,p be offsets. Then $r = q \circ p$ if and only if for every $i \in [1,n]$ such that $i = j+q_j$ and $j = k+p_k$, $r_k = q_j+p_k$.

The next four theorems demonstrate the properties of a group: closure, associativity, and the existence of an identity and an inverse. The sample calculations in Table 2 are helpful in following the proofs. Table 2 also contains a counter-example demonstrating that the group of offsets is non-commutative (non-Abelian), i.e., $p \circ q \neq q \circ p$.

*Theorem 4 (Closure):* Let q,p be offsets. Then $r = q \circ p$ is an offset.

*Proof.* Consider any $i \in [1,n]$. By Definition 3, $k+r_k = k+(p_k+q_j) = (k+p_k)+q_j = j+q_j = i$. By Definition 1, r is an offset.

*Theorem 5 (Associativity):* Let r,q,p be offsets. Then $s = (r \circ q) \circ p = r \circ (q \circ p)$.

*Proof.* Consider any $h \in [1,n]$, and let $h = i+r_i$, $i = j+q_j$, and $j = k+p_k$. By Definition 3, let $qp_k = q_j+p_k$ and $rq_i = r_i+q_j$. Then, $rq_j+p_k = (r_i+q_j)+p_k = r_i+(q_j+p_k) = r_i+qp_k$.

*Theorem 6 (Identity):* Let p be an offset and $z = (0,\ldots,0)$. Then $p \circ z = z \circ p = p$.

*Proof.* Consider any $i \in [1,n]$. To show that $p \circ z = p$, observe that by Definition 3, $i = j+p_j$, $j = k+z_k = k$, and $p_j+z_k = p_k+0 = p_k$. To show that $z \circ p = p$, observe that by Definition 3, $i = j+z_j = j$, $j = k+p_k$, and $z_j+p_k = 0+p_k = p_k$.

*Theorem 7 (Inverse):* Let p be an offset, q be the inverse of p, and $z = (0,\ldots,0)$. Then $q \circ p = p \circ q = z$.

*Proof.* Let $r = q \circ p$, and consider any $i \in [1,n]$. By Definition 3, $i = j+q_j$, $j = k+p_k$, and $r_k = p_k+q_j$. By Definition 2, $r_k = 0$. By Theorem 4(ii), p is the inverse of q, so that by the same argument, $p \circ q = z$.

The guaranteed existence of an inverse parsing formula for each

member of the German to English translator can be used to generate the following parsing steps in an English to German translator. Starting with this English language clause:

, which exhibits the contrast medium applied into the vessels.

[ V D N U P D N ]

, der aufweist das Kontrastmittel applizierte in die Gefaesse.

we apply this parsing sequence:

*Parsandum*

1. [ V D N U P D N ]
2. [ V D N ]
3. [ V ]

*Parsing Formula*

1. 1D6N5 □ U2 □ P3 □ D4 □ N
2. 1[4V2 □ D3 □ N5]
3. (complete)

to obtain:

, which the into the vessels applied contrast medium exhibits.

[ D P D N U N V ]

, der das in die Gefaesse applizierte Kontrastmittel aufweist.

This simple inversion of parsing formulas does not take into account the facts that "exhibits" is an ambiguous noun-verb in English, "contrast medium" is an idiom, and "vessels" has a contextual translation in medicine which differs from its contextual translation, say, in a nautical setting. In other words, a German to English translator cannot simply be inverted to form an English to German translator; but at least there is a group theory inversion principle which allows some of the English to German translation tables to be built up automatically.

## Results and Discussion

The TRANSOFT medical document translator translated the entire computer-readable text of Adler's *Knochenkrankheiten (Bone Diseases)* [1], using translation tables with 30,407 available lexicon entries and 43,945 available parsing formulas. A representative example of the resulting translation is shown in Table 1. The source document contains 7211 sentences, 125,815 words (including start and stop characters for each sentence), 10,217 distinct words, and 859,137 characters. The entire book was translated in 13.0 hours during non-peak periods of computer activity (nights or weekends), an average of 9,671 words translated per hour. This is about 30 times the rate of a human translator. Details of translator performance are given elsewhere [20].

The TRANSOFT system, as well as other practical computer translation systems currently in routine use, employ the design principles of the Russian to English translation system developed at Georgetown University [22]. 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 Organization in Washington, D.C., for translating Spanish to English [12, 13, 22]. 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 [7, 21, 24, 25]. In spite of these limitations, several Georgetown design systems have proven both useful and cost-effective and constitute the majority of computer translation systems in routine use. TRANSOFT appears to be unique in that the computer program itself has a simple design, and its behavior is completely specified by its vocabulary and grammatical rule tables.

As shown in the present report, a TRANSOFT German to English translator can in principle be inverted to form the basis for an English to German translator. This inversion process is limited, however, by the fact that some 20% of lexicon entries are idioms which may not lend themselves well to inversion. Nonetheless, even if only 80% of the translation tables for an English to German translator can be generated automatically, this is a substantial advantage over building an English to German translator de novo. Future development of computer translators will depend upon the wider availability of computer-readable documents, where a draft quality translation can be obtained rapidly without the cost of retyping the document. The investigation of group theory properties, and in particular the inversion property, should assume an important role in translator design and construction.

## REFERENCES


[1] Adler, C.-P.: Knochenkrankheiten. Diagnostik makroskopischer, histologischer und radiologischer Strukturveränderungen des Skeletts. (Stuttgart – New York: Thieme 1983).

[2] Aho, A. V., Ullmann, J. D.: Principles of Compiler Design. (Reading, Mass.: Addison-Wesley 1979).

[3] Arbib, M. A., Manes, E. G.: Monoids and Groups. In M. A. Arbib, E. G. Manes: Arrows, Structures and Functors. The Categorial Imperative. (New York: Academic Press 1975).

[4] Bar-Hillel, B.: Language and Information. Selected Essays on Their Theory and Application. (Reading, Mass.: Addison-Wesley 1964).

[5] Barnett, G. O.: The Application of Computer-based Medical-record Systems in Ambulatory Practice. New Engl. J. Med. *310* (1984) 1643–1650.

[6] Bulkeley, W. M.: Computers Gain as Language Translators Even Though Perfect not They Always. Wall Street J., Febr. 6, 1985, p. 29.

[7] Carbonell, J. G., Cullingford, R. E., Gershman, A. V.: Steps toward Knowledge-based Machine Translation. IEEE Trans. Pattern. Anal. Machin. Intel. PAMI-3 (1981) 376–392.

[8] Chomsky, N.: Aspects of the Theory of Syntax. (Cambridge, Mass.: The Massachusetts Institute of Technology Press 1965).

[9] Garcia-Hidalgo, I., Dunham, G.: An Experiment in English-Spanish Automated Translation of Medical Language Data. Meth. Inform. Med. *20* (1981) 38–46.

[10] Gladkij, A. V., Mel'cuk, I. A.: Elements of Mathematical Statistics. Ed. by J. Lehrberger. (Berlin: Mouton Publ. 1983).

[11] Horowitz, G. L., Bleich, H. L.: Paperchase: A Computer Program to Search the Medical Literature. New Engl. J. Med. *305* (1981) 924–930.

[12] Hutchins, W. J.: Progress in Documentation. Machine Translation and Machine-aided Translation. J. Doc. *34* (1978) 119–159.

[13] Jordan, S. R., Brown, A. F. R., Hutton, F. C.: Computerized Russian Translation at ORNL. J. Amer. Soc. Inform. Sci. *28* (1977) 26–33.

[14] Language and Machines: Computers in Translation and Linguistics. Publication 1416. A Report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C., 1966, pp. 1–34.

[15] Loh, S.-C.: Machine Translation: Past, Present, and Future. ALLC Bull. *4* (1976) 105–114.

[16] Locke, W. N., Booth, A. D.: Historical Introduction. In W. N. Locke, A. D. Booth (Eds): Machine Translation of Languages. (Cambridge, Mass.: The Technology Press of Massachusetts Institute of Technology and New York: Wiley 1955).

[17] Moore, G. W., Hutchins, G. M., Miller, R. E.: Strategies for Searching Medical Natural Language Text: Distribution of Words in the Anatomical Diagnoses of 7000 Autopsied Patients. Amer. J. Path. *115* (1984) 36–41.

[18] Moore, G. W., Miller, R. E., Hutchins, G. M.: Microcomputer Translation for Medical Text: Theorem Verification for Chapter Two of Zeman's Modal Logic. Adv. Math. Comput. Med. (In press).

[19] Moore, G. W., Polacsek, R. A., Erozan, Y. S., de la Monte, S. M., Miller, R. E., Hutchins, G. M., Riede, U. N.: Multilingual Translation Techniques in the Analysis of Narrative Medical Text. Comp. Progr. Biomed. *22* (1986) 35–42.

[20] Moore, G. W., Riede, U. N., Polacsek, R. A., Miller, R. E., Hutchins, G. M.: Automated Translation of German to English Medical Text. Amer. J. Med. (In press).

[21] Schank, R. C., Childers, P. G.: Experiments in Artificial Intelligence. Computerworld *10* (1984) 1–28.

[22] Tucker, A. B., Jr.: A Perspective on Machine Translation: Theory and Practice. Commun. ACM *27* (1984) 322–329.

[23] Watanabe, T., Ohsawa, T., Suzuki, T.: A Simple Database Language for Personal Computers. Commun. ACM *26* (1983) 646–653.

[24] Wilks, Y.: An Artificial Intelligence Approach to Machine Translation. In R. C. Schank, K. M. Colby (Eds): Computer Models of Thought and Language, pp. 114–151. (San Francisco: W. H. Freeman 1973).

[25] Wilks, Y.: A Preferential, Pattern-seeking, Semantics for Natural Language Inference. Artif. Intel. *6* (1975) 53–74.

[26] Wingert, F.: Medical Informatics. (Berlin – Heidelberg – New York – Tokyo: Springer 1981).

[27] Winograd, T.: Computer Software for Working with Language. Sci. Amer. *251* (1984) 130–145.

[28] Winograd, T.: Language as a Cognitive Process. Vol. 1: Syntax. (Reading, Mass.: Addison-Wesley 1983).

Addresses of the authors:
G. William Moore, M.D., Ph.D.,
Grover M. Hutchins, M.D.,
Department of Pathology;

Robert E. Miller, M.D.,
Department of Laboratory Medicine,
The Johns Hopkins Hospital,
600 N. Wolfe Street, Baltimore, MD,
21205 USA;

Richard A. Polacsek, M.D.,
Welch Medical Library, 1900 E.
Monument Street,
Baltimore, MD, 21205 USA;

Urs N. Riede, M.D.,
Department of Pathology,
Ludwig-Aschoff-Haus,
Freiburg University School of Medicine,
Albertstrasse 19,
D-7800 Freiburg,
Federal Republic of Germany.