Translating English into German Using VA File Manager

by Wolfgang Giere and G. William Moore

Abstract

Typically, commercial systems for computer translations are either limited prototypes with a small vocabulary, or high-end systems which require a high volume of source-text to be cost-effective. TRANSOFT is a public-domain computer translator, embedded in the public-domain File Manager of the U.S. Department of Veterans Affairs. TRANSOFT employs grammatical morphology, word and phrase glossaries, ambiguity processing, and word rearrangements. This article describes the successful translation of sample English text from DXplain into German.

Introduction

Computer translators have been discussed in the literature for nearly a half-century, and numerous commercial systems are currently available. [1-9] Typically, commercial systems are either limited prototypes with a small vocabulary, or high-end systems that must process a high volume of source-text to be cost-effective. These commercial systems often do not lend themselves readily to substantial user-specified updates or modifications.

TRANSOFT (TRANslation SOFTware) is a public-domain computer translator, written in the M language, ISO standard 11456. [10,11] TRANSOFT will translate between any two languages using the Latin alphabet. An early version of the M source code is published. [12] TRANSOFT is embedded in the File Manager (FileMan), the core database management and program development environment of the Decentralized Hospital Computer Program (DHCP) of the U.S. Department of Veterans Affairs (VA). [13] The user supplies the dictionary and a grammar in the augmented transition network (ATN) style, which is common to many computer translators. [14] Input is through the File Manager (FileMan) user interface or through an ASCII word processor. The user controls the behavior of the translator through externalized language-specific information and generic program code. TRANSOFT prototype translators have been constructed between English and several languages (German, Dutch, Turkish, romanized Japanese), simply by changing the FileMan databases. [15]

Although it is theoretically possible that this system might be designed and maintained without FileMan and the M language, the project would be considerably more tedious and time-consuming without these powerful tools.

This article examines the use of TRANSOFT to translate an English-language medical text into German. We used a sample text from DXplain, a popular diagnostic decision support system. [16] DXplain was selected as the source document because it: includes English text in all areas of general medicine, is linked to the terminology of the Unified Medical Language System, is available in computer-readable form, is frequently updated, and it serves as a model for other medical systems, especially educational software. [17] The project's success was enhanced by DXplain's relatively unambiguous medical terminology, short sentences, and simple grammatical structures. On the other hand, we took great care to ensure that TRANSOFT properly managed the nuances of German adjectival, noun, and verb inflections.

Why Use VA File Manager?

The VA's DHCP is a hospital information system used in 169 VA medical centers, many government hospitals outside the VA, and in some U.S. private hospitals. [18] There are 85,000 DHCP users worldwide. Parts of the DHCP user interface have been translated and used by hospitals in Germany and Finland. All DHCP source code, lists, and documentation are in the public domain, and may be copied freely.

The FileMan database-management system consists of a set of externally named and internally numbered files. FileMan's user-interface is somewhat cumbersome, but its internal structure is superior to most commercial database-management systems for managing complex lists and dictionaries. FileMan employs either a hierarchical or relational database paradigm, as selected by the user. FileMan served as both a database and a development environment for TRANSOFT. All dictionary, grammar, and source document pre-editing and post-editing functions could be handled by FileMan's database functionality.

TRANSOFT's specific translation operations were carried out by M programs, called as FileMan functions. All searching and pattern-matching in TRANSOFT was performed simply by addressing FileMan global indexes.

TRANSOFT uses four core files for our purposes here: the lexicon file (file 1), the grammar file (file 2), the source document file (file 3), and the target document file (file 4). The lexicon file contains words and phrases translated from the original language (English) into the target language (German); the grammar file contains grammatical paradigms for word rearrangement from the original language into the language to be translated; the source document file contains source documents submitted for translation; and the target document file contains target documents (in German) resulting from translation.

Each file is a rectangular table consisting of fields (=columns) and observations (=rows). The first field is the name field, and every observation has a name. Each field is one of several data types, and consecutive fields are separated from one another internally by the carat (^). The FileMan data array for TRANSOFT is named ^TRS. For example, the 10,415th observation in the lexicon file (FileMan File 1) is:

```
^TRS(1,10415,0)="immediately^immediately^sofort^B".
```

This means that the English word *immediately* (field 1) is translated into the German word *sofort* (field 3), which is an adverb (field 4) (B=adverb, see below). Field 2 is the source word *immediately*, and includes any word-boundary alterations (see below). For an English sentence in the source document which contains *immediately*, the word *sofort* will be placed at a corresponding location in the target document. Any word-order rearrangement between source and target sentences is managed by the TRANSOFT grammar model, using the consecutive parts of speech in the source sentence. The present example is simplified.

The TRANSOFT Grammar Model

TRANSOFT functions by reading each sentence in the source document (English), and translating it into consecutive sentences in the target document (German). Each sentence consists of a sequence of words, and each word or phrase in the source language, including punctuation, is pointed to corresponding word(s) in the target language, and to corresponding parts of speech, using the lexicon file. The sequence of parts of speech (grammar sequence) in the source sentence is called a parsandum, or "thing that must be parsed." The parsandum is pattern-matched to a template in the grammar file, which returns the parts of speech in the stereotypic sequence of the target language. The pattern-match step is nothing more than a global lookup in the appropriate FileMan index global. The ease and simplicity of performing this pattern-match represent a major reason for choosing the M programming language over others. The user customizes the translation by entering one's own word-or-phrase translations, parts of speech, and grammar templates.

The most important design consideration is to determine the parts-of-speech concepts.

The authors spent more than a year, and three hundred meters of fax paper over transatlantic telephone lines, discussing primarily this aspect of the English-to-German model. Each part-of-speech unit is encoded or tokenized by a single Latin letter or selected punctuation mark. A part-of-speech symbol must have corresponding uppercase and lowercase representations, as follows:

Uppercase: ABCDEFGHIJKLMNOPQRSTUVWXYZ[]\/_,
Lowercase: abcdefghijklmnopqrstuvwxyz‹›''-;

That is, *a* is the lowercase for *A*, . . . , and *;* is the lowercase for ,. It is permissible to use traditional parts of speech, or customized parts of speech, as needed. For the present English-to-German grammar model, we included the following parts of speech:

A/=Adjective	P/=Preposition
B=adverB	V/=Verb
D/=Determiner	[=start-sentence
N/=Noun	]=end-sentence

In these parts of speech, for example, an adjective is actually a two-unit concept, consisting of the adjective itself (A) and an appropriate inflection (/), as in *histologisch-e* in the phrase: *der histologisch-e Schnitt*. The forward slash (/) serves as a "grammatical phantom" or place holder for grammatical information, such as gender, number, case inflections, conjugations, etc.

A TRANSOFT word is a sequence of consecutive alphabetical letters occurring in the source document and bounded on either side by blanks or punctuation. A term is a sequence of one or more words and/or punctuation marks, separated by blanks. The number of words in the source and target languages may not have commensurate traditional word boundaries. To establish word equivalence in TRANSOFT, a concatenator (__) glues together contiguous words, a divider (|) breaks a word apart; and phantom word (■) becomes a placeholder. For example, in this English phrase, *Extension through the external sphincter*, the lexicon specifies a target translation and a part-of-speech designator for each word, as follows:

¶ Extension	through	the	external sphincter	. (original phrase)
¶ extension ■	through ■	the ■	sphincter_external ■	. (source terms)
[N /	P /	D /	N /	] (parsandum)
¶ Ausdehnung F	durch A	das ■	Sphincter_externus M	. (target terms)
Ausdehnung	durch	den	Sphincter externus	. (target phrase)

The relevant lexicon elements are as follows:

```
^TRS(1,3554,0)="¶^¶^¶^["
^TRS(1,8582,0) ="extension^extension ■
 ^Ausdehnung F^N/"
^TRS(1,16709,0)="through^through ■^durch A^P/
^TRS(1,16617,0) ="the^the ■^das ■^D /
^TRS(1,8595,0)="external_sphincter
 ^sphincter_external ■^Sphincter_externus M^N/"
^TRS(1,4081,0)=".^.^.^]
```

The part-of-speech sequence, `[N/P/D/N/]`, is the parsandum, or that which must be parsed. The grammar observation for this parsandum is:

```
^TRS(2,2807,0)="[N/P/D/N/]^1[2n0'3p0'4d0'6n7'8]^5;
6;7;9"
```

The argument for this grammar element is `[N/P/D/N/]`, and its value is `1[2n0'3p0'4d0'6n7'8]`. The substrings `1[`, `2n`, `0'`, . . . are particles of the parsing formula. Particle `1[` places ¶ into position 1 in the target sentence; particle `2n` places *Ausdehnung* into position 2 in the target sentence; particle `0'` places the gender marker F (i.e., *Ausdehnung* is a feminine noun) into position 0 (i.e., throws it away) in the target phrase, etc. The resulting order is:

```
1[: ¶                      ¶
2n: extension              Ausdehnung
0[: ■                      F
3p: through                durch
0[: ■                      A
4d: the                    das
0[: ■                      /
5n: external sphincter     Sphincter_externus
0[: ■                      M
6]: .                      .
```

The element `^5;6;7;9` in the parsing formula instructs TRANSOFT to attempt a retranslation of particles *A*, *das/*, and *M*. That is, an accusative-preposition (namely, *durch*) followed by a masculine noun (namely, *Sphincter_externus*) retranslates *das* as *den*. This retranslation results in the German phrase, *Ausdehnung durch den Sphincter externus*.

Successful Results

A partial copy of the English-language DXplain text file, a total of three hundred disease nodes, was uploaded into the source document file, and translated entirely into German. The lexicon file contained 713,893 bytes (6,604 FileMan entries), the grammar file contained 237,861 bytes (2,079 File-

Man entries), and the source document contained 347,680 bytes (5,339 FileMan entries). The source document contained 5,339 sentences, and a total of 21,746 terms, 6,604 of them distinct. The frequency distribution of the 40 most-frequent terms in the lexicon is shown in table 1. The 45 (0.7%) most frequent terms represented one-quarter of all terms, and the 436 (6.6%) most frequent terms represented half of all terms. The translation required a total of 5,339 grammar formulas, 2,079 of them distinct. As with the word distribution, there was a steep downward trend in grammar-formula frequency. That is, the 32 (1.5%) most frequent formulas represented half of all formulas in the sample text. Sample translations by this system are shown in figure 1. A reader skilled in German will note that there are a few minor errors in the German target text, which can be corrected either by further refinements in the translator or by post-editing the target text. In particular, ambiguous prepositions denoted by an asterisk (*), such as *in*bei*, *mit*bei*, *von*durch*, *zu*bis*, require post-editing by people.

Example 1. Bacterial infection; drainage of nose impaired by engorged nasal mucosa as in allergic, viral rhinitis; aggravated during air travel.
Translation: Bakterielle Infektion; Nasendränage gestört durch Verstopfung mit Nasenschleim wie in allergischer, viraler Rhinitis; verschlimmert bei Flugreisen.
Example 2. Tenderness on pressure over frontal or maxillary sinus; postnasal discharge. Transillumination: reduced, absent illumination of affected sinus.
Translation: Druckempfindlichkeit der Stirn- oder Kiefer-Höhle; postnasaler Ausfluß. Transillumination: reduzierte, fehlende Beleuchtung des befallenen Sinus.
Example 3. Salt-losing adrenal hyperplasia; defect in hydroxylation of c-21; failure of 21-hydroxysteroids.
Translation: Elektrolytverlust bei Nebennierenhyperplasie; defekte Hydroxylierung des c-21; Versagen der 21-Hydroxysteroide.
Example 4. Genetic defect in activity of enzyme homogentisic acid oxidase leading to accumulation, excessive excretion of homogentisic acid.
Translation: Genetische defekte Aktivität der Enzym Homogentisinsäure führend zu Anhäufung, exzessiver Exkretion der Homogentisinsäure.
Example 5. Arthritis; stiffness, pain in large joints, spine in 50 percent of patients; alkalosis; bluish-black pigmentation of cartilaginous structures as sclera, pinna, nose, tendons, nails.
Translation: Arthritis; Steifheit, Schmerz in*bei großen Gelenke, Wirbelsäule in*bei 50 Prozent der Patienten; Alkalose; blauschwarze Pigmentierung der Knorpelstrukturen als*wie Sklera, Außenohr, Nase, Sehen, Nägel.

Figure 1. Sample English-to-German translations.

DXplain as a Useful Translator

The high cost and relative inflexibility of commercial translation systems have kept them out of reach of many investigators who might profitably use them. There are many translation jobs for which the limited vocabulary and cumbersome word-entry procedures of low-end translators renders these translators effectively useless; but the expense of high-end translators cannot be justified. This article exemplifies such a translation job: DXplain contains a large text file with a highly specialized vocabulary not available in inexpensive translators. The DXplain computer-readable text file is continually updated, which recognizes that a one-time translation rapidly becomes out of date. On the other hand, the text file requires no separate data-entry step and very little editing beforehand for computer translation; the sentences are typically short and unambiguous; and the vocabulary, while highly specialized, is often small, with relatively few irregular forms. The initial TRANSOFT translation was slightly more expensive than hiring a professional human translator, but we expect future retranslations to be much cheaper.

We assert that, as with words, grammar formulas sufficient to translate most of the document belong to a short list, a sort of Zipf's Law for grammar formulas.

Even in this short sample text, the statistical behavior of the text appears to satisfy an important property. There is a short list of forty-five terms which account for one-quarter of all term occurrences. These are common words in any English-language medical document, and many of the words are common in almost any English-language document (table 1). Thus a lexicon containing this short list of words is already sufficient to translate a large proportion of words in the document. This extremely high frequency of a few words may be characterized by Zipf's Law. [19] If one determines the frequency of each word in the document, and sorts the words in the descending order of frequency, then the most frequent word has rank 1, the second-most-frequent word has rank 2, etc. Zipf's Law states that word frequency is inversely proportional to word rank. On linear graph paper, the plot of word frequency against word rank is a hyperbola; on log-log graph paper, the plot of word frequency against word rank is a straight line with a negative slope. A graph showing the logarithm of word rank versus the logarithm of word frequency for the DXplain source document is shown in figure 2, and demonstrates a near-linear relationship between these variables, as predicted by Zipf's Law. Large samples of English, German, and Chinese text have been shown to satisfy Zipf's Law. [20,21,22] The essential feature of Zipf's Law is that a short list of words accounts for a large proportion of all word occurrences in any natural language text.

Rank	Frequency	English	German
1	1431	of	■ G das ■
2	626	in	in*bei D das ■
3	469	possibly	möglicherweise
4	325	with	mit*bei D das ■
5	273	prognosis	Prognose F
6	264	or	oder
7	124	usually	normalerweise
8	117	by	von*durch D
9	91	as	als*wie
10	86	and	und
11	78	increased	erhöht ■
12	77	to	zu*bis D das ■
13	75	from	von*durch D das ■
14	71	normal	normal ■
15	61	for	für A das ■
16	58	unknown	unbekannt ■
17	53	after	nach D das ■
18	53	infection	Infektion F
19	51	small	klein ■
20	51	type	Typ MS
21	48	pulmonary	pulmonal ■
22	46	associated with	begleitet von D
23	46	especially	besonders
24	45	blood	Blut SES
25	45	chronic	chronisch ■
26	44	during	während G das ■
27	44	features	Anzeichen P
28	44	large	groß ■
29	43	guardedly favorable	vorsichtig günstig ■
30	43	on	auf D das ■
31	43	rarely	selten
32	40	cells	Zellen P
33	40	form	Form F
34	39	edema	Ödem SS
35	39	favorable	günstig ■
36	39	onset	Ausbruch MS
37	37	tumor	Tumor MS
38	37	variable	variabl ■
39	35	absence of	Abwesenheit*Fehlen F von D
40	35	hereditary	erblich ■

Table 1. Forty most frequent words in DXplain text sample.

Similarly, there is a short list of thirty-two grammar formulas, which account for half of all grammar-formula occurrences in the sample text. We assert that, as with words, grammar formulas sufficient to translate most of the document belong to a short list, a sort of Zipf's Law for grammar formulas. A graph showing the logarithm of grammar-formula rank versus the logarithm of grammar-formula frequency for the DXplain source document are shown in figure 3, and again demonstrate a near-linear relationship between these variables.

The consistent presence of Zipf's Law as a property of both words and grammar formulas in natural-language text suggests a strategy for constructing both the lexicon and grammar. First, one obtains a frequency distribution for both words and parsing formulas. Then one should fashion the major design considerations around these high-frequency elements in the lexicon or grammar. Next, moderate-frequency elements should be fashioned around the models created for high-frequency forms. Finally, low-frequency forms that do not fit the existing design may be flagged for prior editing or post-editing. We have used this approach to guide our models for prepositions followed by an ambiguous case (dative or accusative), gerunds (English verb forms ending in -ing), and separable verb forms (e.g., *stop* may translate as *auf . . . hören*).

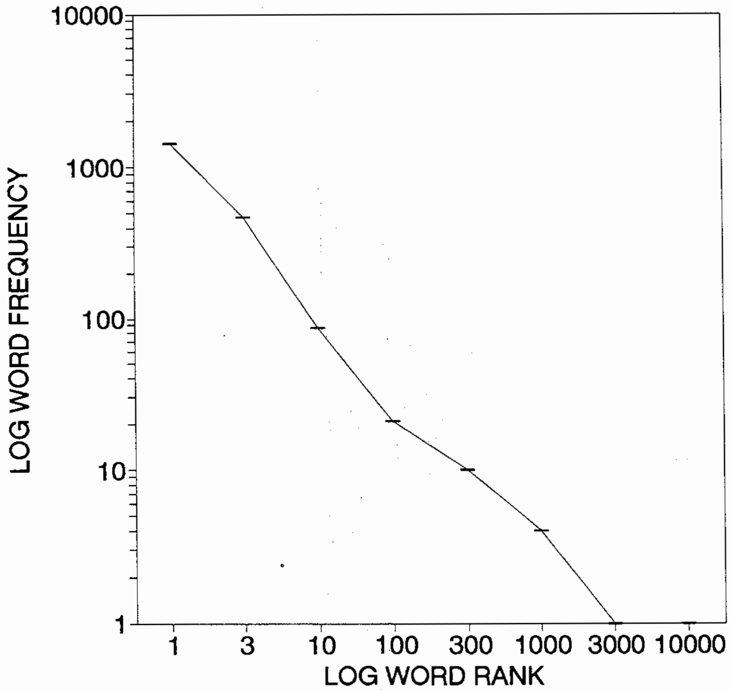

Figure 2. Logarithm of word rank (abscissa) versus logarithm of word frequency (ordinate) for the DXplain source document. In a frequency distribution of words, the most-frequent word has rank 1, the second-most-frequent word has rank 2, etc. Zipf's Law states that word frequency is inversely proportional to word rank, so that on a log-log plot, one expects a negative slope, linear relationship between these variables, as shown.

Semantic ambiguities require a variety of methods for disambiguation, including a thesaurus or world model, such as the *Medical Subject Headings* (MeSH) or *Befunddokumentation und Arztbriefschreibung* (BAIK). [23,24] We expect to apply our methods to the entire DXplain, and to other, computer-readable medical texts.

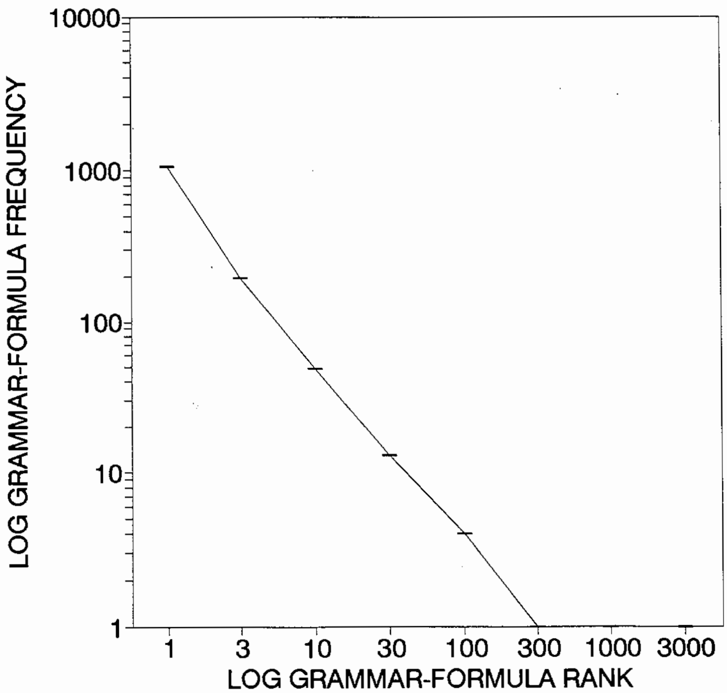

Figure 3. Logarithm of grammar-formula rank (abscissa) versus logarithm of grammar-formula frequency (ordinate) for the DXplain source document. In a frequency distribution of grammar formulas, the most-frequent formula has rank 1, the second-most-frequent formula has rank 2, etc. Zipf's Law applied to grammar formulas would predict that grammar-formula frequency is inversely proportional to grammar-formula rank, so that on a log-log plot, one expects a negative slope, linear relationship between these variables.

Eventually, VA FileMan will be used to construct a translator's workstation. TRANSOFT can also be used for encoding free-text diagnoses or for translating stereotypic free text into PROLOG statements. [25,26]

The Affordability Factor

The DXplain experiment exemplifies an ideal application for TRANSOFT. First, the source document was prepared initially in computer-readable form, so that there was no preparatory cost for the source document. Otherwise, and particularly if the document's typist is not conversant with the source language, then the cost of producing an error-free source document for translation alone is probably prohibitive. Second, the DXplain source document consisted predominantly of short sentences, written in a small vocabulary, and in a simple grammatical style. Third, we were able to exercise limited editorial control over spelling and punctuation. Again, this control presupposes an editor who is conversant with the language of the source document. Thus one cannot readily import source documents in unknown source languages; one is limited to broadcasting target documents which are translated from one's own source language. Finally, we expect to retranslate DXplain on a periodic basis. Thus the cost of preparing the lexicon and grammar is spread over many translations.

We were dismayed at the absence of computer-readable lexicons available for a project like this. Commercially available bilingual dictionaries cost ten times (or more) the price of a comparable paper dictionary. These dictionaries typically have a small vocabulary and an awkward user-interface, which cannot be customized easily. In fact, the computer dictionaries are more time-consuming to use than a paper dictionary; we can't understand why one would pay a premium price for inferior functionality. Unless this commercial market improves, TRANSOFT users must reckon with the cost of preparing their own dictionaries and grammars for the foreseeable future. Perhaps there must be a *samizdat* or "underground" dictionary in circulation for a few years before commercial publishers venture into this field.

In conclusion, we have translated selected text from DXplain, a book-sized computer-readable source document, from English into German. The 5,339 sentences required a lexicon of 6,604 distinct words and a grammar of 2,079 sentence templates. The study suggests that TRANSOFT can translate computer-readable documents with a small vocabulary, simple grammatical style, and which are likely to be revised and retranslated on a periodic basis. For these source documents, translation quality will be adequate, and the effort of dictionary and grammar preparation can be spread over multiple translation jobs. TRANSOFT can make computer translation available to many researchers who otherwise could not afford it.

Prof. G. Octo Barnett, M.D., provided the sample English text from DXplain and Katharina Nimmo assisted in preparing the lexicon.

Wolfgang Giere, M.D., is professor and chairman at the Center of Medical Informatics, J.W. Goethe University Medical Center, Frankfurt, Germany. Prof. Giere is a member of the M Development Committee; editor of *M Börse* (M Marketplace); chairman of MUG-D (Germany); and a cofounder, former chairman, former executive director, and currently honorary member of MUG-Europe.

G. William Moore, M.D., Ph.D., is a staff pathologist and chief of the autopsy section at the Baltimore VA Medical Center, Baltimore, Maryland, and holds appointments in the pathology departments at the University of Maryland and The Johns Hopkins University. He has been working with M Technology since the late 1970s.

See endnotes on page 23.

Endnotes

1. W.N. Locke and A.D. Booth, "Historical Introduction," in *Machine Translation of Languages* (New York: Technology Press of MIT and John Wiley & Sons, 1955).
2. M. Vasconcellos, "Overview: Machine Translation," *BYTE* 18 (1993): 152–166.
3. E. Hovy, "How MT Works," *BYTE* 18 (1993): 167–176.
4. "Resource Guide: Machine-translation Software," *BYTE* 18 (1993): 185–186.
5. M. King, "Are There any Lessons to be Learned from Machine Translation?" in *Computerized Natural Medical Language Processing for Knowledge Representation*, ed. J.R. Scherrer, R.A. Côté, and S.H. Mandil, (Amsterdam: Elsevier Science Publishers B.V., 1989), 73–82.
6. S. Nirenburg, ed., *Machine Translation: Theoretical and Methodological Issues*, (London: Cambridge University Press, 1987).
7. W.J. Hutchins, *Machine Translation: Past, Present, Future*, (Chichester: Ellis Horwood Ltd., 1986).
8. M. Vasconcellos and M. Leon, "SPANAM and ENGSPAN: Machine Translation at the Pan American Health Organization," *Comput Linguist* 11 (1985): 122–136.
9. A. Tucker, "A Perspective on Machine Translation: Theory and Practice," *CACM*, 27 (1984): 322–329.
10. G.W. Moore et al., "TRANSOFT: Medical Translation Expert System," *Artif Intell Med* 1 (1989): 149–157.
11. G.W. Moore et al., "Automated Translation of German to English Medical Text, *Am J Med* 81 (1986): 103–111.
12. G.W. Moore, R.E. Miller, and G.M. Hutchins, "Microcomputer Translator for Medical Text: Theorem Verification for Chapter Two of Zeman's Modal Logic," *Adv Math Comput Med* 7(1986): 1621–1633.
13. R.G. Davis, *FileMan: A User Manual*, (Bethesda, Maryland: National Association of VA Physicians, 1987).
14. W. Woods, "Transition Network Grammars for Natural Language Analysis," *CACM*, 13 (1970): 591–606.
15. G. W. Moore, "TRANSOFT."
16. G.O. Barnett et al., "DXplain—an Evolving Diagnostic Decision Support System," *Journal of the American Medical Association*, 258 (1987): 67–74.
17. R.G. Davis, *FileMan*.
18. U.S. Department of Health and Human Services, National Library of Medicine: Unified Medical Language System Knowledge Sources, 4th exptl ed. (Bethesda, Maryland: National Library of Medicine, 1993).
19. G.K. Zipf, "On the Economy of Words," in *Human Behavior and the Principle of Least Effort* (Cambridge, Massachusetts: Addison-Wesley Press, Inc., 1949), 19.
20. G.W. Moore et al., "Integrated Pathology Reporting, Indexing, and Retrieval System Using Natural Language Diagnoses," *Modern Pathol* 1 (1988): 44–50.
21. W. Giere, "Foundations of Clinical Data Automation in Cooperative Programs," in *Fifth Annual Symposium of Computer Applications in Medical Care*, ed. H.G. Heffernan (Washington, D.C.: IEEE Computer Society Press, 1981), 1142–1148.
22. Q. Zhang, "Easy Entry of Chinese Character Set Symbols," in *Fifth Annual Symposium of Computer Applications in Medical Care*, 143–149.
23. G.W., Moore, "TRANSOFT."
24. W. Giere, *BAIK: Befunddokumentation und Arztbriefschreibung im Krankenhaus* (Taunusstein: Media, 1986).
25. G.W. Moore and J.J. Berman, "Performance Analysis of Manual and Automated SNOMED Coding," *Am J Clin Pathol* (1993, on press).
26. W. Giere and I. Wakai, "TRANSPRO: Natural Language to PROLOG Translation of Genealogy Statments in USDVA File Manager," *Artif Intell Med*, 3 (1991): 139–147.