Meta
Journal des traducteurs
Translators' Journal

META

The Translation of S.N.O.P.: A First Step toward the Construction of an Automated Medical Lexicon

Bruce Barkman, Lise Bernier, Léo Cousineau et Gabrielle Tanguay

Volume 19, numéro 1, mars 1974

La traduction médicale

URI : https://id.erudit.org/iderudit/001974ar
DOI : https://doi.org/10.7202/001974ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Barkman, B., Bernier, L., Cousineau, L. & Tanguay, G. (1974). The Translation of S.N.O.P.: A First Step toward the Construction of an Automated Medical Lexicon. *Meta*, *19*(1), 28–42. https://doi.org/10.7202/001974ar

The Translation of S.N.O.P.: A First Step toward the Construction of an Automated Medical Lexicon

Medical terminology provides an exemplary illustration of rapid language change. The World Health Organization has estimated that several thousand new terms are created annually. Many of these creations are the products of new discoveries in the biomedical sciences, as in the field of genetics. Others result from theoretical reorientations, as in the reclassification and renaming of viruses. Many others are duplications or synonyms of existing medical concepts. For a single medical entity, there may be as many as 30 synonyms [1]. A more serious problem than neologisms or the proliferation of synonyms is the use of a single term to designate different medical concepts. For example, *petit mal* may account for either 3% or 80% of all cases of epilepsy, depending on how the specialist groups the different types of seizures involved. This type of homonymy may cause confusion across language boundaries as well as within the same language. Thus « schizophrenia », « chronic bronchitis » and « peptic ulcer » have different meanings in German, French and English, formal linguistic similarities across the three languages notwithstanding [2].

Various national and international groups have attempted to standardize medical terminology, such as the Comité d'étude des termes de médecine in Canada and the World Health Organization. Their admirable efforts have not reached the vast majority of medical personnel, however, and many special fields of biomedicine have not yet confronted the problems involved in the development of a standardized systematic terminology.

The most recent large-scale effort to record medical usage and standardize medical terminology in French was begun in 1956 and reached fruition in 1972, with the appearance of the third volume of the *Dictionnaire français de médecine et de biologie* [3]. This reference work, prepared under the direction of Dr. Alexandre

1. A. Manuila, L. Manuila, M. Nicole and H. Lambert, *Dictionnaire français de médecine et de biologie,* 3 vol., Paris, Masson, 1970-1973, p. VII; vol. 4 : in press.
2. *Ibid.*, p. IX-X.
3. *Ibid.*

Manuila, editor-in-chief, contains approximately 150 000 medical terms and definitions. A fourth volume will include various indexes and cross references. We will discuss this dictionary in more detail further on, since it has been an indispensable tool in our translation efforts. The fact that 16 years have elapsed since the inception of the project and its successful completion is no indicator of editorial indolence, but simply a sign of the tremendous scope and difficulty of medical terminology and usage.

At Sherbrooke, we are trying to develop a medical lexicon which can be used for a variety of purposes, including : 1. the codification of medical diagnoses for reportage of health statistics to provincial, federal and international bodies ; and 2. the compilation of coded medical information for basic and applied research, teaching and patient care [4]. To meet these goals, we decided that the lexicon should have certain characteristics usually absent from ordinary classificatory systems or dictionaries.

First, it should be possible to delete and add entries at any time, so that updating can be accomplished without the time-lags involved in normal publications. Revised editions of the *International Classification of Diseases, Adapted (I.C.D.A.)* [5] for example, appear at 10 year intervals, so that a given edition is likely to be out-of-date medically almost as soon as it appears.

Second, we wanted every natural language expression in the lexicon to be assigned a code, each part of which would have a particular semantic value in the field of medicine. This would mean that instead of arbitrary assignment of code numbers which have no semantic significance, each code would provide, because of its hierarchically ordered structure, a concise definition of the linguistic expression associated with it. Such a coding system would enable us to achieve the equivalent of the defining sections of dictionaries and to avoid the enormous amounts of space required by definitions expressed in ordinary language. Compared to other fields, the semantic networks of medicine are few in number, and so we thought that the development of an adequately structured coding system would not prove too intractable. In fact, such a code had already been devised by the American College of Pathologists, in the *Systematized Nomenclature of Pathology (S.N.O.P.)* (Chicago, 1965), and we decided to use it as the basis for the lexicon.

Third, we wanted the lexicon to be bilingual at first, and eventually multilingual. We considered that if we could find equivalent linguistic expressions across languages for the highly structured medical concepts in the lexicon, it would be a significant step toward the standardization and internationalization of medical terminology. And to take a more practical viewpoint, a French-English lexicon which permitted the coded parts of a patient's record to be retrieved in either language would facilitate the transfer and use of the record from one hospital to another, or even from one country to another.

Finally, we wanted to computerize the system ; an automated system seemed the only realistic approach, if the features described above were to be realized.

4. This work is supported by Health and Welfare Canada, Grant Number 605-21-33.
5. World Health Organization, 8th rev. ed., 1965.

Coding of diagnoses is usually done manually by medical record librarians, but if the coding can be automated, the librarians, who have considerable medical training, can be freed for other duties. Further, automatic coding would make it possible to access the various parts of the record as they are generated, thus providing instant care capacities. At present, the patient's record is created manually, with resultant delays between the time that the record is written and the time that certain parts of it can be used in the delivery of health care. Manual translation of the patient's record would result in further delays, and the volume of health records virtually precludes this possibility. For hospitalized patients alone, there were 2 404 728 separations from Canadian hospitals and 881 971 from Quebec hospitals in 1969 [6]. And the number of records for hospitalized patients is considerably smaller than the total number of medical records, since a record has to be opened for every member of the population who visits a doctor. Also, since most translators lack training in the biomedical sciences, the possibility of errors creeping into the patient's record through translation was a disturbing prospect. So we decided that we would attempt to effect automated translation of medical diagnoses, using French and English versions of *S.N.O.P.* as a point of departure. In passing, we should note that we are not concerned with such prerequisites of good translations as elegance of expression and fidelity to the stylistic aspects of the original texts, but only with their medical content.

The scope of the project, then, is rather large. Seen in its entirety, it involves the computerization of the patient's health record, using a multilingual semantically well-ordered lexicon as the basis for the processing and coding of the medical information contained in the record. Various retrieval techniques will ensure the usefulness of the computerized records in patient care, teaching, research and the compilation of health statistics.

At present, we are attempting to test the feasibility of such a system by limiting ourselves to the domain of pathology, for a number of reasons. First, pathology includes enough aspects of clinical medicine to provide a sufficiently complicated testing ground for our hypotheses. The extension from pathology to clinical medicine requires the addition of only one semantic category to *S.N.O.P.* Second, a computerized English version of the *Systematized Nomenclature of Pathology* exists already at the National Institutes of Health in Bethesda, Maryland, along with an encoder which permits accurate automatic coding of English pathology diagnoses [7]. We have created a French version of the nomenclature and are working on the French linguistic rules for the encoder so that French diagnoses can also be coded by the computer system at N.I.H.[8] Similar strategies are being followed for a German version of *S.N.O.P.*, and if the French and German encoding efforts are successful, we will be well on the way toward the creation of a multilingual medical lexicon.

6. *Hospital Morbidity*, Statistics Canada, Health and Welfare Division, 1969, p. 9.
7. A.W. Pratt, « Automatic Processing of Pathology Data », *Journées d'information médicale*, Toulouse (France), Institut de recherches d'informatique et d'automatique, 1971, p. 595-609.
8. B. Barkman, « Linguistics and the Construction of an Automated Medical Lexicon », paper delivered at the 31st Annual Convention of the Canadian Association of Medical Record Librarians, September 7, 1973.

S.N.O.P. contains approximately 15 000 medical expressions, each of which has been placed into one of four semantic categories. The first category is *Topography*. Each item in the topography section refers to a particular anatomical entity or body site, for example, « tracheal cartilage ». The second category, *Morphology,* contains items which describe structural alterations, including traumatic abnormalities such as « ultraviolet burn » and neoplasms such as « adenosquamous carcinoma ». The third semantic division, *Etiology,* provides an ordered list of the causes of abnormalities, including pathogenic organisms, chemicals and drugs and physical agents. Examples for each are « smallpox virus », « ethyl phosphate » and « automobile ». The fourth category, *Function,* lists the signs and symptoms of diseases and biochemical and enzyme disorders, such as « postural syncope » and « galactose disorder ».

FIGURE 1

Illustration of hierarchical ordering of S.N.O.P. codes

T 2601 Mucous membrane of bronchus

T = Topography
2 = Respiratory tract
6 = Bronchus
1 = Mucous membrane

Every linguistic expression in *S.N.O.P.* is assigned a five character code which uniquely identifies it. Furthermore, the code is hierarchically ordered so that each character contributes to the definition of the linguistic expression. In Figure 1, the « T » places the item in the semantic field of topography, the « 2 » indicates that the item is in the respiratory tract, and the « 6 » and the « 1 » show progressively finer subdivisions of the respiratory tract. Most items in the nomenclature have similarly ordered codes. For example, every item in the respiratory tract begins with « T 2 », such as « T 2441 Right vocal cord » and « T 2701 Mucous membrane of bronchiole ».

These four semantic fields provide an adequate framework for the analysis of pathology diagnoses, since every diagnosis contains a combination of medical terms, almost always nouns or noun phrases, each of which fits into one of the four fields. For instance, *athéromatose de l'artère fémorale* describes a morphological abnormality, *athéromatose* (M 5212) in a body site, *l'artère fémorale* (T 4740). Some of the diagnoses we are using as test data are surgical pathology reports from the Saint-Vincent-de-Paul Hospital of Sherbrooke. These reports typically contain an item from the morphology section and one from topography. Diagnoses which contain items from etiology and function are quite rare in surgical pathology reports. However, we are also using autopsy reports from the Centre hospitalier universitaire de Sherbrooke, and these contain items from all categories. *Abcès de la paroi abdominale* (Escherichia coli) shows an abnormality, *abcès* (M 4174), in a body site, *paroi abdominale* (T Y430), caused by an etiologic agent, « escherichia coli » (E 1341). The following diagnostic statement, *glomérulonephrite chronique avec insuffisance rénale,* contains an item from the

function category, *insuffisance* (F 9005). This expression also illustrates how linguistic entities smaller than a word can provide semantic information which affect the assignment of *S.N.O.P.* codes, thus necessitating morphological analysis for successful code assignments. « Glomerulo- » is assigned to the code T 7120 *glomérule, nephr-* to T 7100 *rein* and *-ite* to M 4001 « inflammation ». Since the inflammation is chronic, the exact code would be M 4301, where the « 3 » indicates « chronic ». Further on, we will discuss some of the linguistic rules wereby language expressions which are acceptable in pathology diagnoses but are not in the *S.N.O.P.* dictionary are transformed into *S.N.O.P.* terms.

While most *S.N.O.P.* codes are associated with only one linguistic expression, the structure of the nomenclature also allows the inclusion of several expressions under the same code. This occurs when there are *synonyms, equivalent terms* or *eponyms* for a particular concept. « Synonym » has not been defined by the editors of *S.N.O.P.,* but we assume that if one linguistic expression is semantically identical to another, they are synonymous. In this case, the first linguistic expression listed is called the preferred term, and expressions indented under it are synonyms, as illustrated in Figure 2. No criteria are mentioned for distinguishing preferred terms from synonyms, but the choice of preferred terms seems to depend on frequency of use estimations by the editors.

FIGURE 2

Preferred and synonymous terms

T 2200	Sinus accessoire SP (sans précision)	*Preferred term*
	Sinus de la face	*Synonyms*
	Sinus accessoire du nez	

Equivalent terms *are not synonyms,* but represent groupings of *related but distinct medical concepts,* all of which are assigned to the same code. In the first example of Figure 3, all the terms are considered equivalent because « ... they represent cutaneous manifestations of a tuberculous process elsewhere in the body [9] ». The second example groups various parts of the hippocampus under the same code number.

FIGURE 3

Equivalent terms in S.N.O.P.

M 4475	Tuberculide SP	*Preferred term*
	Erythème induré	
	Tuberculide papulo-nécrotique	*Equivalent terms*
	Tuberculide rosacéiforme	
T X257	Hippocampe	*Preferred term*
	Alvéus	
	Circonvolution godronnée	*Equivalent terms*
	Sillon fimbriogodronné	

9. *Systematized Nomenclature of Pathology,* College of American Pathologists, Chicago (Ill.), 1965, p. XVII.

A medical expression which contains the name of its discoverer or of someone closely identified with it is called an eponym. An example is *tumeur de Grawitz,* the eponym for *adénocarcinome à cellules claires, hypernéphrome* and *carcinome à cellules rénales.* In *S.N.O.P.,* eponyms are usually treated as synonymous terms, unless the eponym is the only expression for the medical concept, as in *corps de Herring* or unless the eponym is the most used term, as in *trompe de Fallope* for *trompe utérine.* In both cases, the eponym is considered the preferred term. The number of eponyms in medical terminology is staggering and biomedical personnel are attempting to purge most of them from active use. This is partly because there are often several eponyms for a single medical concept. *Syndrome du scalène antérieur* has the following eponyms : *syndrome de Nafziger, syndrome du défilé costo-claviculaire de Leriche, syndrome de Coote, syndrome de Coote-Hunauld, syndrome de Haven, syndrome de Nonne* and *syndrome d'Adson.* Not all these would be known to a particular doctor, and he would be unable to access all the relevant literature pertaining to this syndrome under only one of the eponyms. Further, and more important, a proper name provides no medical information. In the preceding example, *scalène antérieur* and *défilé costo-claviculaire* give some notion of the medical concepts involved, but the others do not. For these reasons, attempts are being made to use medically descriptive terms instead of eponyms.

An English version of *S.N.O.P.* is stored on disc at the Division of Computer Research and Technology at N.I.H., as is a preliminary French version. We have completed the revisions and corrections to the French version and are entering them via terminal. The final French version of *S.N.O.P.* should be available on disc by the end of October, 1973.

The N.I.H. encoder uses the disc-stored version of *S.N.O.P.* as a matching dictionary for texts from pathology reports. The basic technique used is a lookup matching process, whereby items from the pathology reports are checked against relevant sets from the disc-stored *S.N.O.P.* The encoder works in several phases. The first phase requires the transformation of the language of the pathology diagnoses into the data structures that the encoder operates on. Parts of speech labels may be assigned here and specification of the various linguistic rules which may be applied to individual items are attached as well, along with restrictions on the application of these rules. The second phase, or lookup matching process, uses the data resulting from the first phase to « ... fetch from disc, those *S.N.O.P.* dictionary entries relevant to the encoding of an utterance [10] ». The final phase compares the revised input texts and matches them against the texts of the appropriate *S.N.O.P.* entries, and prints out the codes and the *S.N.O.P.* preferred term associated with each code, provided that exact matches are found. Redundant matches are deleted here. The entire encoding process, which involves many linguistic paraphrastic rules, is designed to transform the text of the original diagnoses into the language of *S.N.O.P.* entries, so that automatic coding of the

10. G. Dunham, *Pathology Diagnoses Language Encoder,* Internal Report, D.C.R.T., National Institutes of Health, Bethesda (Md.), 1971, p. 9.

diagnoses can be effected. These paraphrase rules can be considered a kind of intra-language translation.

The N.I.H. encoder is very expensive to operate and can be used only with very large computer installations, such as the I.B.M.-370's at N.I.H. This state of affairs results partly from using *S.N.O.P.* entries as the items to be matched. At Sherbrooke, we think that the construction of a stem and affix dictionary would reduce the number of necessary lookups considerably, especially if it is combined with auxilliary stores containing syntactic and semantic information. We think that appropriate *S.N.O.P.* codes can be generated without first obtaining an exact match of the original pathology texts against the natural language texts of the *S.N.O.P.* dictionary entries. Experiments using a stem-affix dictionary developed for an extended German *S.N.O.P.* developed by Drs. Friederich Wingert and Peter Graepel are in progress at N.I.H., and promising results have already been achieved. For instance, the entire German stem-affix dictionary, consisting of 25 000 medical expressions, can be read by the computer in 14 seconds. We intend to use the N.I.H. encoder at first, on our French test data, but we are also planning the development of our own encoder, which we hope will prove more economical to operate.

As mentioned previously, we are restricting our lexicon to the domain of pathology for the moment. The *S.N.O.P.* Committee is attempting to extend and revise *S.N.O.P.* so that it will be capable of dealing with clinical medicine. The nomenclature will be called *S.N.O.M.E.D.*, or the Systematized Nomenclature of Medicine [11]. The extension of *S.N.O.P.* to clinical medicine requires the addition of a new semantic category, *Procedures*, which will include operations, examinations, tests, and administration of drugs. A trial edition of *S.N.O.M.E.D.* will be tested internationally, beginning in January, 1973. After the trial edition has been revised, we hope to be able to provide a French translation and extend our automatic coding efforts to *S.N.O.M.E.D.* Since a typical patient's record would necessitate at least 16 five character codes, manual coding would be extremely time-consuming and automated data-processing techniques are virtually imperative.

When the first French translation of *S.N.O.P.* was begun, the only high-quality bilingual dictionary of medical terminology available was the long out-of-print *Dictionnaire français-anglais/anglais-français des termes médicaux et biologiques* [12]. This work contains approximately 56 000 entries in all, or about 23 000 for each language. As with most bilingual dictionaries, an entry is usually defined by its translation only, and the definitions of single-language dictionaries are generally absent. Since this dictionary was published in 1952, developments in medicine have made it, naturally enough, an unreliable tool for current medical terminology. The structures of D.N.A. and R.N.A. were unknown for example, and the first translation of « pacemaker » was *nœud sinusal, nœud de Keith et Flack* and the second was « pacemaker (*cœur*) ». Further, the artificial pacemaker, or *stimulateur cardiaque*, was not listed at all. Enzyme disorders, such as

11. The *S.N.O.P.-S.N.O.M.E.D.* Committee meets quarterly under the direction of Dr. Roger A. Côté, Centre hospitalier universitaire de Sherbrooke.
12. P. Lépine, Paris, Flammarion, 1952.

	WEBIN	66	1-		V

o implant (living tissue) so as to form an organic union (as in a lesion) /

	EXPRE	73	36		

acile de —— à un patient un fragment de sa propre peau prélevé sur une autre du corps. /26 mars/

	eml

META XIX, 1

plastic transplantation	DOMED	69			

ansplantation of tissue between individuals belonging to different species.

greffe	EXPRE	73	38		

reffe d'un fragment de peau ou d'un organe d'un animal sur un autre animal. rs/

	eml

META XIX, 1

graft	TAMED	69			N

——skin or other living substance inserted into a similar substance to supply an absence or defect by attachment and growth into an integral part of the original substances.

greffon	EXPRE	73	36		

et, parfois l'impossible se réalise : peu à peu, le greffé et le —— apprennent à se supporter. /26 mars/

NCB	eml

META XIX, 1

transplant patient	TIMEC	73	49		

/ that disease /bone-marrow malignancy/ occurs about 100 times more frequently in —— than it does in members of the general population / /March 19/

greffé	EXPRE	73	36		N

et, parfois, l'impossible se réalise : peu à peu, le —— et le greffon apprennent à se supporter. /26 mars/

NCB	eml

META XIX, 1

ly	TIMEC	73	49		

stence of antibodies — agents produced by the body in response to the challenge ase-causing organisms — was discovered / /March 19/

ps	MAMED	70	1-		

ubstance spécifique de nature globulinique apparaissant dans le sérum ou les tissus te de l'action d'éléments exogènes ou endogènes / appelés antigènes.

IFR	eml

META XIX, 1

ous graft; **ft**	DOMED	69			

graft of tissue derived from another site in or on the body of the organism receiving it. utologous graft, p. 628/

effe	EXPRE	73	38		

reffe d'un fragment de peau ou d'un organe d'un animal sur le même animal. ars/

	eml

META XIX, 1

autoimmune disease	MIMED	72			

——disease due to immunologic action of one's own cells or antibodies on components of the body.

maladie auto-immune	EXPRE	73	86		

quand l'organisme commence / à attaquer certains tissus de son propre corps, c'est ce qu'en termes techniques on nomme les ——, immunes contre soi. /5-11 mars/

NHD NFQ	eml

META XIX, 1

autoplastic transplant	DOMED	69			

——the transplantation of tissue between different parts of the same individual.

autogreffe	EXPRE	73	38		

——greffe d'un fragment de peau ou d'un organe d'un animal sur le même animal. /26 mars/

NCB	eml

META XIX, 1

	DOMED	69		G	

a stored supply of human material or tissues for future use by other individuals /

ıe d'organes	EXPRE	73	39	S	

que des méthodes de conservation par le froid existent et permettent même de rendre
e des embryons de souris / il est désormais envisageable de constituer de vraies——.
ars/

NCB	eml

META XIX, 1

e agent	TIMEC	73	48		

sts began to suspect that the body had a mechanism for identifying and combatting
only after Louis Pasteur discovered the existence of bacteria / /March 19/

oathogène	MAMED	70	1-		

gent (chimique, physique, mécanique ou biologique) qui provoque une maladie.

	eml

META XIX, 1

chemotherapy; drug treatment	TIMEC	73	48		

the older techniques — surgery, radiation and —— (——) — have been used successfully in bringing some cancers under control. /March 19/

chimiothérapie	MAMED	70	1-		

——/ administration d'un produit chimique spécifique afin de guérir une maladie cliniquement reconnaissable ou d'enrayer sa progression.

NID	eml

META XIX, 1

foreign cell	TIMEC	73	51		

thus lymphocytes, which know their body's own cells, recognize others as foreign and trigger an immunological alarm. /March 19/

cellule étrangère	EXPRE	73	39		

les lymphocytes T reconnaissent à sa forme l'antigène qui caractérise la ——. /26 mars/

NFR NHD	eml

META XIX, 1

« cholesterol esterase disorder », were also missing. Moreover, the Lépine dictionary was intermediate in size between pocket and unabridged of full-scale dictionaries, and so many medical terms were absent, even though they were known at the time. For example, Dorland's dictionary [13] contains approximately 86 000 entries of English medical terms, whereas the 19th edition of the Garnier and Delamare dictionary [14] contains only about 14 000 entries, an increase of about 1 000 terms over the 18th edition. Despite its limitations, the Lépine dictionary was the basic tool of reference for the first translation, simply because no better bilingual work existed which attempted to cover all of medicine. Used in conjunction with the unilingual dictionaries just mentioned as well as medical textbooks, plus consultation with bilingual medical personnel, a satisfactory French list of the preferred terms in the *S.N.O.P.* dictionary was prepared by Sr. Tanguay. The translation was limited to preferred terms partly because the reference works used did not list synonyms and equivalents very frequently, but also because it was hoped that French medical usage would prove more uniform than English, so that only the preferred terms would occur in pathology diagnoses. Also, with rare exceptions, the doctors consulted did not offer alternate expressions for the medical concepts of their specialities. This is undoubtedly because it was difficult enough for them to find exact French equivalents for the English terms, and also because their own linguistic practice made them unable to think of several French expressions for a given medical concept, and so they produced only the terms which they themselves were accustomed to use. Thus, although we were aware of paraphrastic possibilities, we wanted to avoid their inclusion in French *S.N.O.P.*

The hope that French medical usage would prove more uniform than English proved a vain one, not surprisingly to those versed in the study of language, for the possibilities of paraphrase are probably equally likely for all languages and even for subsets of languages as restricted semantically as medical French and English. To mention only two examples from our surgical pathology test data, *polype simple d'une corde vocale* is equivalent to *nodule du chantre,* and *endomètre à la phase secrétoire* is medically the same as *endomètre à la phase lutéinique.* Therefore, we were forced to add the synonyms and equivalent expressions to the French version of *S.N.O.P.,* so that we would be able to use the *N.I.H.* encoder successfully. How we were to distinguish all of the sets of equivalent and synonymous French expressions with our limited resources remained problematic, and for a time we set this translation task aside to work on the morphological analysis and transformational rules of the preferred terms.

Luckily, the third volume of the Manuila dictionary became available at the beginning of 1973, so we acquired the entire alphabetic listings of this important work (*A* through *M* in two volumes had been available since 1971). At last we had a French dictionary of medicine which was up-to-date and comprehensive enough to compare more than favorably with Dorland's. (In fact, we think it surpasses Dorland's in the quality of its definitions.)

13. *Dorland's Illustrated Medical Dictionary,* 24th ed., Philadelphia, W.B. Saunders, 1965.
14. *Dictionnaire des termes techniques de médecine,* Saint-Hyacinthe (Québec), S.O.M.A.B.E.C., 1972.

If a medical concept has more than one linguistic label, the editors of the Manuila dictionary have defined it under the expression recommended by a national or international authority or under the expression that is most often used. Thus, *bourgeon du goût,* in the opinion of the editors, is the preferred term for certain kinds of taste buds, and its synonyms are *oignon du goût* and *corpuscule du goût. Indicateur isotopique* is recommended by A.F.N.O.R. (Association française de normalisation) and its synonym is *traceur isotopique.* The basic approach taken by the editors to the problem of synonyms is that most of them are unnecessary or incorrect, and various usage labels are employed to indicate this. For *stimulateur cardiaque* the synonym « pacemaker » is given with the following warning : « *(néologisme d'origine anglo-américaine, couramment employé dans les textes français, déconseillé. Il désigne d'ailleurs également un centre anatomique d'automatisme cardiaque)* ». Nonetheless, all synonyms are given separate entries, with cross references to the relevant preferred term, to aid in bibliographic searches and in the identification of unknown terms. If a user does not know the expression *branche cutanée dorsale de la main* but looks for *rameau dorsal de la main* he will be referred to the first term, where the definition occurs.

Another good feature of the dictionary is the identification of the originator of a term in the linguistic notes section of the entry. For *capacité affinitaire,* the linguistic note provides the information that it is a « *terme créé par Tiffeneau* ». In the fourth volume, there will be indexes of proper names, key words and cross references of other types. Etymological information has also been relegated to this volume. Since words of Greek and Latin origin account for about 75% of all medical terminology, much space has been saved in the body of the dictionary, and innumerable repetitions of the same information avoided, by this editorial device. The final volume of this work will also contain a guide to word-creation for researchers in the biomedical sciences, in the hope of encouraging them to invent linguistically sound and scientifically descriptive neologisms for the concepts and findings of their endeavors [15].

Using our now provisional list of preferred French terms, we were able to check our translations against the entries in the Manuila dictionary, to make the necessary revisions, and to find a whole range of synonyms and equivalent terms whose medical accuracy has already been ascertained. To arrive at our final French translation of *S.N.O.P.,* we used the following basic procedure : 1. we took each provisional French equivalent and checked it against the Manuila entry ; 2. we chose as the preferred term the principal entry in Manuila ; 3. we listed possible synonyms ; 4. we submitted the translations and the original English terms to the director of the project, Dr. Léo Cousineau, or to a bilingual specialist in a particular area [16] ; 5. if we had made a medical error, the translation was revised ; 6. if North American French usage differed from the preferred term listed in Manuila, we chose the term used most often in Quebec, as reflected in

15. A. Manuila, L. Manuila, M. Nicole and H. Lambert, *Dictionnaire français de médecine et de biologie,* p. XXIII-XLV.
16. We are particularly grateful to Dr. P. Dionne of Saint-Vincent-de-Paul Hospital, Sherbrooke, for his advice on surgical terms and to Dr. G. Dupuis of the Centre hospitalier universitaire de Sherbrooke for his corrections of the biochemistry sections.

our test corpus of surgical and autopsy pathology reports. Although the Manuila dictionary reflects current international opinion about medicine and medical usage of the French language, we were constrained by the structure of the N.I.H. encoder to choose as the preferred term the linguistic expression which occurs most frequently in pathology texts. Actually, this problem did not arise too often. One example is *hypernéphrome,* the preferred term in Manuila. We chose *adénocarcinome à cellules claires* as our preferred term because it occurs more frequently than *hypernéphrome* in our test data. This choice will facilitate the matching and lookup devices of the encoder. This example provides some evidence for the notion that medical practice in Quebec reflects North American, rather than European norms, since the English preferred term is « clear cell adenocarcinoma ». Similar cases abound in North American French medical terminology. The linguistic similarity is undoubtedly influenced by the fact that many French-speaking doctors receive some of their training in English-speaking environments and subsequently use calquing as the most common technique of translating the concepts they have learned into French.

As we proceeded with our translation, we noticed a great many patterns of linguistic similarity between English and French equivalent terms. Among others, we have found the following correspondances for suffixes listed in Figure 4. Naturally, it is not the case that a single suffix in one language always corresponds to a single suffix in the other, as a glance at the English possibilities for French *-ie* will show. A considerable number of the suffixes bear semantic as well as grammatical information. For instance, in addition to identifying the items as nouns, the suffixes *-oma / -ome* almost always indicate the semantic marker « tumor », and *-itis / -ite,* « inflammation ». « *-Osis* is a noun-forming suffix which indicates

FIGURE 4

English-French suffix correspondances

English	French	Examples
-itis	-ite	phlebitis / phlébite
-osis	-ose	salmonellosis / salmonellose
-oma	-ome / -oma	carcinoma / carcinome
-ism	-isme	pleomorphism / pléomorphisme
-ectomy	-ectomie	hysterectomy / hystérectomie
-tomy	-tomie	tracheotomy / trachéotomie
-ix	-ice	appendix / appendice
-ium	-e	endometrium / endomètre
-iasis	-iase	lithiasis / lithiase
-ia	-ie	anemia / anémie
-y	-ie	anomaly / anomalie
-ism	-ie	embolism / embolie
-ity	-ité	irritability / irritabilité
-ous	-eux, -euse	fibrous / fibreux

that the noun belongs in the function category of *S.N.O.P.* as in *brucellosis,* or else in the morphology category, as in *necrosis* ». The semantic marker is « abnormal condition ». In Figure 4, the roots or stems to which the suffixes are attached are identical for both languages. (We have ignored French accents, since our computer printouts are always in capital letters.) This is not always the case, of course, but usually only minor spelling adjustments are necessary, as in *personality / personnalité.*

In addition to single words which have identical forms in both languages except for the suffix, and those which are totally different, such as *kidney* and *rein,* we found many equivalent single words that were orthographically identical or partially so in both languages. Some identical forms are *muscle, fracture, radiation* (almost all words ending in *-ation* are identical in both languages), and many chemical compounds and elements such as *cobalt* and *phosphate.* A common pattern of partial similarity is absence of a final *-e* for the English word as in *lymph / lymphe, branch / branche, acid / acide, gland / glande,* and so on.

Some partially similar equivalents which occur for more than one pair of tokens are : *viscera / viscère, vulva / vulve ; neurosis / névrose, neuralgia / névralgie* and *mesentery / mésentère, artery / artère.*

There are also many derivational paradigms which correspond in both languages. A derivational paradigm is a set of derivational affixes which occur with a given set of stems. Figure 5 shows some partial derivational paradigms for French and English equivalent terms. The formal linguistic similarities dovetail with semantic similarities for these derivational paradigms. Thus *-oid / oïde* means « resembling », *-osis / -ose* indicates an abnormal condition and *-oma / -ome* indicates a tumor. It is interesting to note that these formal and semantic correspondances between English and French medical terms also exist for the medical vocabularies of other Indo-European languages, a reflection of the common Greek and Latin origins of medical terminology in all of them. These resemblances make the task of automated translation somewhat easier for the domain of medicine than for the semantically and syntactically far less restricted domains of political science or literature.

FIGURE 5

Partial derivational paradigms in French and English

English	French
-IC ; -ITIS ; -OMA	-IQUE ; -ITE ; -OME
cystic cystitis cystoma	cystique cystite cystome
hepatic hepatitis hepatoma	hépatique hépatite hépatome
keratic keratitis keratoma	kératique kératite kératome
-OID ; -OSIS ; -OMA	-OÏDE ; -OSE ; -OME
sarcoid sarcosis sarcoma	sarcoïde sarcose sarcome
lymphoid lymphosis lymphoma	lymphoïde lymphose lymphome
carcinoid carcinosis carcinoma	carcinoïde carcinose carcinome
myeloid myelosis myeloma	myéloïde myélose myélome

For single words with more than one root, we found that the order of the roots within the word remained the same in the majority of cases, as in *pancreatoduodenal / pancréatoduodénale* or *cystadenoma / cystadénome*. But the latter example presented some problems. If this expression is unmodified, Manuila gives *cystadénome* as the preferred term. However, if it is modified by *hépatique,* the roots are reversed in the preferred expression, giving *adénocystome hépatique,* although *cystadénome hépatique* is a synonym with no usage restrictions. « Hepatic cystadenoma » does not exist in *S.N.O.P.,* because the site of a tumor is taken from the topography section. Thus there would be two codes, M 8440 for « cystadenoma » and T 5600 for « liver ». For cases like this, we decided to keep the root order the same in both languages, especially since the editorial choice of preferred term over synonym seems quite arbitrary in the first place, and there would be no *S.N.O.P.* entry for *adénocystome hépatique* as such, because the structure of *S.N.O.P.* requires the separation of body sites and morphological alterations.

When it is necessary to translate sequences longer than a single word, it is a truism that interlanguage cognates are not always appropriate. For instance, although « surface » is a word in both French and English, the translation of « placental fetal surface » is *face fœtale du placenta,* and « effect of internal reduction » is appropriately rendered *résultats de réduction fermée,* and not *effet de réduction interne.* In English pathology, it is appropriate to speak of « pleural fluids », but in French pathology « fluids » is generally rendered *liquides.* Interestingly enough, most of the non-cognate terms indicate very general semantic categories, such as the ones just mentioned, as well as « system » (where French prefers *appareil* most often) and « disorder » where French uses the more specific *déficience* for enzyme disorders). In the language of pathology, we have found that if the terms are quite specific semantically, then the use of English-French cognates results in French preferred terms in the vast majority of cases. The two major exceptions we found to this rule-of-thumb are « pituitary gland » and « nitrogen », where the preferred terms in French are *hypophyse* and *azote* respectively, rather than the cognate equivalents *glande pituitaire* and *nitrogène.*

An orthographic convention for English *S.N.O.P.* expressions consisting of more than one word is the use of commas and word order changes to indicate various syntactic and semantic relationships. French pathology expressions do not allow this linguistic technique, so we had to reinterpret the English expressions before attempting translations. Sometimes the comma simply means that a modifier has been moved from its normal position before the noun. « *Ulcer, healed* », « *ankylosis, fibrous* » and « *placenta, twin* » are illustrations of this use of the comma, and the translations *ulcère guéri, ankylose fibreuse* and *placenta gemellaire* presented no particular problems. Another use of the comma in English involves the elimination of certain function words as in « postoperative state, implantation » and « fibrosis, septal, liver ». In the first illustration, an implantation is an operation which requires the placement of a substance into the body, such as radium. The comma replaces a preposition or phrase such as

« as a result of » or « after ». In French, the appropriate meaning could be arrived at with an expression like *état postopératoire à la suite d'une implantation* but current usage prefers the expression *état postopératoire avec implantation,* which is the translation we decided upon, although we are fully aware of the semantic ambiguities of this phrase in the context of French as a whole. In the domain of French pathology, however, ambiguity does not arise, so we sacrificed linguistic precision to the customary usage of medical personnel. The second expression means fibrosis in the septal portions of the liver, where « septal » is not a kind of fibrosis, but identifies certain parts of the liver, even though the normal phrase, without commas, would be « septal fibrosis of liver ». We have translated this concept as *fibrose du septum hépatique,* thus attempting to preserve the semantic relationships in its formal realization. We have made similar decisions whenever no conflicts existed between ordinary usage and precision of expression. Our experience has been that attempts to legislate language usage are almost always unsuccessful, no matter how desirable such legislation might be on logical, cultural or practical grounds. This may be a defeatist position, but it reflects the following reality : pathologists are busy people and their summary diagnoses for a linguistic code which communicates semantic information adequately enough through a fairly small set of stereotyped expressions ; they are not going to change their linguistic formulas to suit language reformers. We do not encourage sloppiness of expression, on the contrary. Whenever a choice existed among several commonly used linguistic expressions, we have chosen the one which communicates the medical concept most clearly, as in the fibrosis example. But where no choices among equally current expressions existed, we took a seemingly contradictory position, as in the case of the postoperative states.

Although we have had a considerable number of problems in translating *S.N.O.P.* from English to French, some of which are discussed above, the fact that struck us the most was the existence of regular linguistic correspondances between equivalent medical concepts. These regularities are the result of a largely cognate vocabulary and the semantic restrictions imposed by the nature of the discipline of pathology. A few examples of regular patterns of correspondance for phrases are *yolk stalk / tige vitelline* (French use of an adjective for the English noun), *Hassall's corpuscle / corpuscule de Hassall* (the regular possessive construction for both), *closed fracture / fracture fermée* (past participles and normal word order in both), and *nuclear-cytoplasmic ratio a'teration / modification de la proportion nucléaire cytoplasmique* (typical English preference for multiple-word premodification structures and corresponding French preference for postmodifiers introduced by prepositions). Many others could be cited, although of course no single correspondance is universally applicable. Needless to say, the occasional lacks of regular correspondances kept us alert and added spice to what might otherwise have proved a tedious task.

As we mentioned earlier, the N.I.H. encoder includes components which transform the language of pathology diagnoses into the medically equivalent, but linguistically different, expressions contained in the disc-stored version of *S.N.O.P.*

We call these components the *metalanguage,* and they contain various sorts of linguistic rules, or more precisely a context-sensitive set of paraphrastic rules, which effect a kind of intra-language translation. To arrive at an accurate match and the correct *S.N.O.P.* codes for the diagnosis *urétrite prostatique ulcéreuse,* the encoder has to identify the topographic site as T 7511 *urètre prostatique* and the morphological alteration as M 4003 *inflammation ulcéreuse.* After *urétrite prostatique* has been recognized as a significant syntactic unit, using the principle of longest match, the encoder identifies the *-ite* suffix and assigns the code 3nte to *urétrite,* which means that the last three characters are to be cut off, an *E* added to what remains and that the transformed noun is in the topography category. The « -ite » is transformed into « inflammation » by a morphosemantic conversion rule, and grouped with *ulcéreuse.* The French *S.N.O.P.* dictionary is then consulted, appropriate matches for the transformed phrases are found, and the final printout of the diagnosis is given as follows :

T 7511 *urètre prostatique*

M 4003 *inflammation ulcéreuse*

The diagnosis can also be printed in English, if this is desired, by using the code for English dictionary entries. (The fourth column of each entry identifies the language of the entry : *1* identifies an English language expression, *2,* a French one, *3,* German.) This routine has to be applied after the *S.N.O.P.* codes have been assigned, however, because translation from the texts has not yet been attempted. The present procedure to obtain printouts in languages other than the original thus involves a table lookup routine.

We are currently working on the establishment of the entire set of rules for the French metalanguage and hope to have a trial version ready for testing at the end of 1973. Eventually, we hope to generate the codes directly from the original texts rather than from the matching-lookup devices. But this will become possible only when the code-generating capabilities and the hierarchical ordering of *S.N.O.P.-S.N.O.M.E.D.* have been revised and refined further.

In conclusion, we would like to offer a few reassuring words to those translators who may still fear, after reading this article, that we are trying to put them out of business with our automated analysis and transformation of natural language texts. First, we are far from being able to approximate a good human translation of even the restricted area of pathology diagnoses. The best we can do, with many reservations, is to code the medical information in the diagnoses and provide a kind of shorthand linguistic clue to the original text, a text which lacked most of the syntactic resources of a natural language in the first place. Second, as mentioned above, human translators could not possibly manage the enormous numbers of medical diagnoses produced annually. Third, the nature

of automatic processing of natural language data requires extremely precise information about the rules which govern language, without which the processing devices simply cannot be successful. Thus our investigations can only help to make specific something that good translators have always known — how languages work [17].

BRUCE BARKMAN, LISE BERNIER,
LÉO COUSINEAU et GABRIELLE TANGUAY

17. We would like to acknowledge the indispensable aid of our colleagues at N.I.H., Dr. A.W. Pratt, Dr. M. Pacak, and Messrs. M. Epstein, G. Dunham and W. White.