A Medical Language Processor for Two Indo-European Languages

Ngo Thanh Nhan[1,3], Naomi Sager[1], Margaret Lyman[2], Leo J. Tick[2],
François Borst[3], Yun Su[4]

[1] Courant Institute of Mathematical Sciences, New York University, NY, NY
10012, USA; [2] NYU Medical Center, NY,NY 10016, USA; [3] Division of Informatics, University
Hospital of Geneva and Faculty of Medicine, CH-1211 Geneva 4, Switzerland;
[4] State Economics Information Center, 58 Sanlihe Road, Beijing, China

ABSTRACT

The syntax and semantics of clinical narrative across Indo-European languages are quite similar, making it possible to envison a single medical language processor that can be adapted for different European languages. The Linguistic String Project of New York University is continuing the development of its Medical Language Processor in this direction. The paper describes how the processor operates on English and French.

A. INTRODUCTION

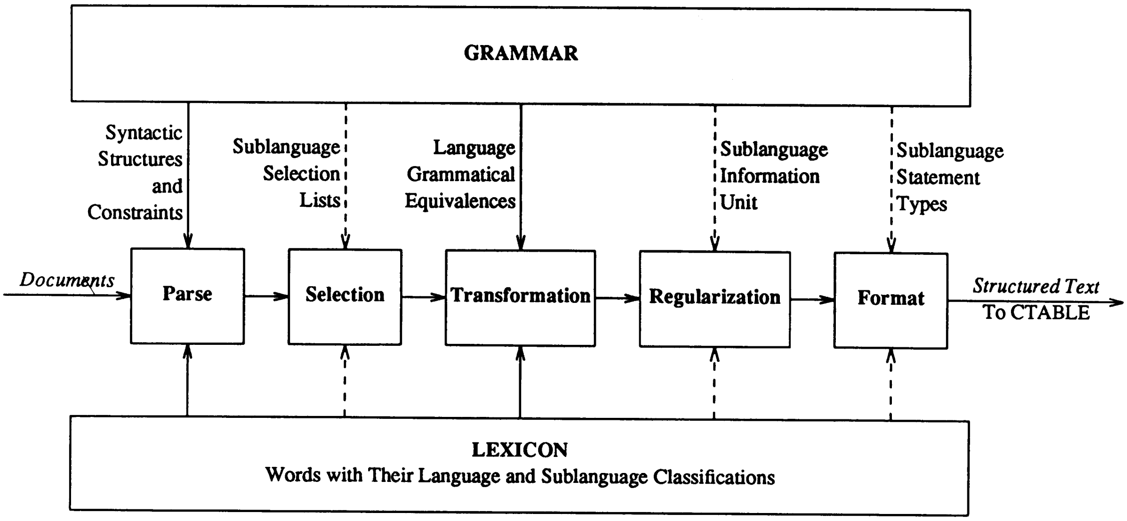
Is it possible to organize the information in clinical narrative algorithmically? Yes, if the algorithm is based on the principles by which language carries information. One can treat language as a code: a very complex, sometimes ambiguous code, but one that is not -- like many artificial codes -- designed to hide the message. Human ("natural") language is, despite its potential for vagueness and ambiguity, an effective mechanism for transmitting information, and it does so by utilizing its own structural properties as the "code" [1].

Using this approach, the Linguistic String Project (LSP) of New York University has developed a system that converts textual information from its linear form to an explicitly structured form. First it uses grammatical relations to determine the gross structure of successive sentences; then (in part simultaneously) it uses the regularities of language usage that are characteristic of the applicational area to refine, label, and finally rearrange the linguistic/informational units of the discourse into a database of semanticaly organized textual information.

Our main area of application has been the narrative of patient records, i.e. the "sublanguage" [2] of clinical reporting. Progress in the development of this Medical Language Processor for use on English-language hospital discharge summaries and ambulatory visit reports has been reported in previous SCAMC and MEDINFO volumes [3,4] as well as in book form [5]. Currently, we are adapting the system for French *Lettres de Sortie* in a joint project with the Hopital Cantonal Universitaire de Genève [6,7].

A companion paper [8] presents retrieval results from French documents. The present paper illustrates how structural similarity and a common sublanguage have made the adaptation of a Medical Language Processor from one Indo-European language to another not unduly difficult.

Figure 1
LSP Medical Language Processor


GRAMMAR
Syntactic Structures and Constraints
Sublanguage Selection Lists
Language Grammatical Equivalences
Sublanguage Information Unit
Sublanguage Statement Types
Documents
Parse
Selection
Transformation
Regularization
Format
Structured Text
To CTABLE
LEXICON
Words with Their Language and Sublanguage Classifications

0195-4210/89/0000/0554$01.00 © 1989 SCAMC, Inc.

B. MEDICAL LANGUAGE PROCESSING VIA LSP

The LSP System component grammars are modules which execute sequentially and are compiled cumulatively [Figure 1]. Each component of the grammar consists of BNF definitions of parse tree structures, procedures that operate on these structures (routines, restrictions and transformations) and lists which state the well-formed parse tree combination of medical word classes.

To illustrate the processing, we will take an example from an outpatient visit report of a pediatric patient:

> 'Was seen in emergency room 2 days
> ago for diaper rash and given bacitracin
> and oral antibiotic.'

Translated into French [by a French physician], it reads:

> 'A été vu en salle d'urgence il y a 2
> jours pour un érythème fessier et a reçu
> de la bacitracine et un antibiotique oral.'

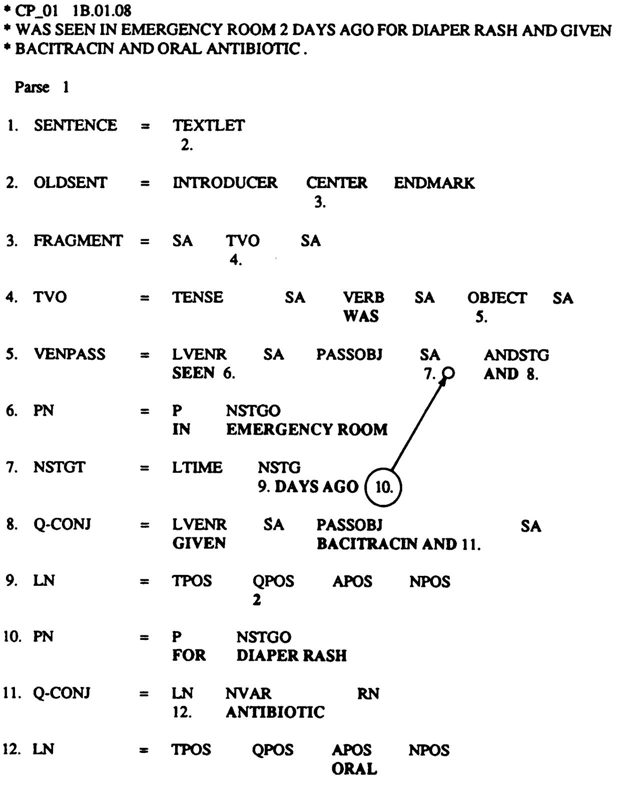
B.1. PARSING GRAMMAR

Figure 2 shows an output of parsing the example English sentence using the LSP English string grammar [9], and Figure 3 shows the corresponding French parse, obtained using the same grammar adapted for French. Internally, the outputs are trees; the short form of output seen in Figures 2 and 3 displays the segments ("strings") of the decomposition, and the position at which a modifier occurs, to the right or left of its host word, as indicated by the placement of the number corresponding to the line number where the modifier string is written. It will be seen that the component strings are much the same in the two decompositions. A difference is that English has more reduction around conjunction (leaving out a second *'was'* before *'given'*) and we have treated *'diaper rash'* as a single lexical item. Also French supplies more articles. The result is that the English output has fewer lines.

Notice that the English parse, Figure 2, line 7 (and the French parse, Figure 3, line 8) contains an incorrect modifier placement of the prepositional phrase PN *'for diaper rash'* (Fr. *'pour un érythème fessier'*) as modifier of the time phrase *'2 days ago'* (Fr. *'il y a 2 jours'*). This parse is syntactically possible (by the BNF definitions) but impossible in the medical sublanguage. The arrow pointing from 10. to the right of 7. in Figure 2 (and from 15. to the right of 11. in Figure 3) suggests the correct adjunction. The next component adjusts the parse accordingly.

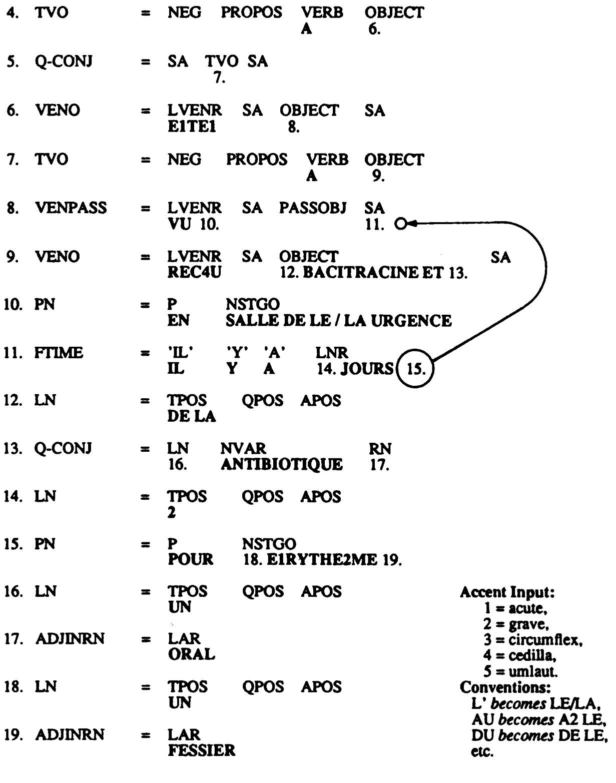
Surprisingly few changes to the English grammar were needed to accommodate French. Although the dictionaries that give parts of

Figure 2
English parse



```
* CP_01   1B.01.08
* WAS SEEN IN EMERGENCY ROOM 2 DAYS AGO FOR DIAPER RASH AND GIVEN
* BACITRACIN AND ORAL ANTIBIOTIC .

  Parse  1

 1.  SENTENCE   =  TEXTLET
                     2.

 2.  OLDSENT    =  INTRODUCER   CENTER    ENDMARK
                                  3.

 3.  FRAGMENT   =  SA      TVO       SA
                            4.

 4.  TVO        =  TENSE          SA      VERB     SA      OBJECT    SA
                                          WAS                5.

 5.  VENPASS    =  LVENR      SA      PASSOBJ      SA        ANDSTG
                   SEEN  6.                        7. O      AND  8.

 6.  PN         =  P        NSTGO
                   IN       EMERGENCY ROOM

 7.  NSTGT      =  LTIME       NSTG
                               9. DAYS AGO (10.)

 8.  Q-CONJ     =  LVENR      SA       PASSOBJ                    SA
                   GIVEN               BACITRACIN AND 11.

 9.  LN         =  TPOS        QPOS       APOS       NPOS
                               2

10.  PN         =  P          NSTGO
                   FOR        DIAPER RASH

11.  Q-CONJ     =  LN         NVAR             RN
                   12.        ANTIBIOTIC

12.  LN         =  TPOS       QPOS       APOS       NPOS
                                         ORAL
```

Figure 3
French parse



```
* CP_01  1B.01.08
* A E1TE1 VU EN SALLE DE LE / LA URGENCE IL Y A 2 JOURS POUR UN
* E1RYTHE2M2 FESSIER ET A REC4U DE LA BACITRACINE ET UN
* ANTIBIOTIQUE ORAL .

 Parse  1

 1.  SENTENCE    =  TEXTLET
                      2.

 2.  OLDSENT     =  INTRODUCER    CENTER   ENDMARK
                                    3.

 3.  FRAGMENT    =  SA      TVO     SA     ANDSTG
                             4.            ET  5.

 4.  TVO         =  NEG    PROPOS   VERB    OBJECT
                                     A        6.

 5.  Q-CONJ      =  SA   TVO  SA
                          7.

 6.  VENO        =  LVENR    SA   OBJECT     SA
                    E1TE1           8.

 7.  TVO         =  NEG     PROPOS   VERB   OBJECT
                                      A       9.

 8.  VENPASS     =  LVENR    SA   PASSOBJ   SA
                    VU  10.                 11.  O<-

 9.  VENO        =  LVENR    SA   OBJECT                           SA
                    REC4U         12. BACITRACINE ET 13.

10.  PN          =  P        NSTGO
                    EN       SALLE DE LE / LA URGENCE

11.  FTIME       =  'IL'      'Y'   'A'     LNR
                    IL         Y     A      14. JOURS (15.)

12.  LN          =  TPOS      QPOS   APOS
                    DE LA

13.  Q-CONJ      =  LN      NVAR             RN
                    16.     ANTIBIOTIQUE     17.

14.  LN          =  TPOS      QPOS   APOS
                    2

15.  PN          =  P         NSTGO
                    POUR      18. E1RYTHE2ME 19.

16.  LN          =  TPOS      QPOS   APOS                     Accent Input:
                    UN                                            1 = acute,
                                                                  2 = grave,
17.  ADJINRN     =  LAR                                           3 = circumflex,
                    ORAL                                          4 = cedilla,
                                                                  5 = umlaut.
18.  LN          =  TPOS      QPOS   APOS                     Conventions:
                    UN                                          L' becomes LE/LA,
                                                                AU becomes A2 LE,
19.  ADJINRN     =  LAR                                         DU becomes DE LE,
                    FESSIER                                     etc.
```

speech and word subclassifications are necessarily different, the lexical categories, both grammatical and medical, are largely the same.

B.2. SELECTION GRAMMAR

The selection component's main job is to check the well-formedness of sublanguage word class combinations in a sentence, in terms of cooccurrence lists of adjective and noun clusters (LIST N-ADJ), noun and noun clusters (LIST N-NPOS), clusters of prepositional phrase and its host (LIST P-NSTGO-HOST), and variations of subject-verb-object clusters (LIST S-V-O or BE-S-O). For example, the selection component recognizes the ill-formed combination of noun and prepositional modifier *'days'+'for'+'diaper rash'* (*'days'* NTIME1 + *'for'* + *'diaper rash'* H-INDIC) of the parse in Figure 2 and proceeds to correct this by using a prepositional phrase selection list P-NSTGO-HOST [Figure 4]. Similarly, the component uses the list of well-formed medical sublanguage combination for *'pour'* to correct the parse tree of the French sentence.

The P-NSTGO-HOST list asserts that the cooccurrence requirements for a preposition *'for'* with its object *'diaper rash'* H-INDIC is one of the well-formed medical subclass combinations, but NUNIT and NTIME1 (attributes of *'days'*) are not in the list of combinations with *'for'*+H-INDIC. The selection component then looks for a new modifier location for *'for diaper rash'* (and Fr. *'pour un érythème fessier'*): as a sentence modifer of *'was seen in emergency room'* (Fr. *'a été vu en salle d'urgence'*).

The selection component does not reject a parse received from the parsing component, but assigns node attribute FAIL-SEL to phrases which do not match any cooccurrence patterns. If a cluster passes a cooccurrence pattern, a node attribute SELECT-ATT or ADVERBIAL-TYPE (for specific phrases such as TIME-ADVERBIAL, ADJUNCT-TYPE, CONN-TYPE, BODYLOC-PN, QUANT-ADVERBIAL, INSTR-TYPE) will be assigned to the node with the value chosen [Figure 4].

Figure 4

LIST P-NSTGO-HOST

Combination of Prepositional Phrase and Syntactic Host

ENGLISH VERSION			
Preposition	*Word Class of Object NSTG*	*Phrase Type*	*Word Class of Syntactic HOST*
'for'	H-INDIC	NO-TYPE	H-TTCOMP, H-TTGEN, H-TTCHIR, H-TTMED, H-TXCLIN, H-TXSPEC, H-TXPROC.
	H-DIAG	NO-TYPE	H-TTCOMP, H-TTGEN, H-TTCHIR, H-TTMED, H-TXCLIN, H-TXPROC, H-TXSPEC, H-RECORD.
	H-PT	ADJUNCT-TYPE	H-INST, H-DOCTOR, H-TTMED, H-TTCOMP, H-TTGEN, H-RECORD.
	H-PTAREA	BODYLOC-PN	
	H-PTPART	BODYLOC-PN	
	H-TMBEG	TIME-ADVERBIAL	
	H-TMPER	TIME-ADVERBIAL	
	NTIME1	TIME-ADVERBIAL	
	NUNIT	QUANT-ADVERBIAL	

FRENCH VERSION			
Preposition	*Word Class of Object NSTG*	*Phrase Type*	*Word Class of Syntactic HOST*
'pour'	H-TTCHIR H-INDIC		H-TTGEN H-TTCOMP,H-TXVAR

B.3. TRANSFORMATION COMPONENT

The aim of this component is to normalize the sentence into ASSERTIONs or FRAGMENTs corresponding to chunks of related information in the target FORMATs (to be discussed later). It first fills in the information gaps due to conjunction ellipsis, turns imperative and interrogative sentence types into affirmative, reunifies verbal splits (due to past or present participle), and expands relative clauses into full assertions.

The example sentences actually each consist of three assertions conjoined by *'and'* (Fr. *'et'*). So that the sentence can be described as below, with ___ marking showing gaps caused by conjunction ellipsis.

'Was seen in emergency room
2 days ago for diaper rash
and ___ given bacitracin
and ___ ___ oral antibiotic.'

'A été vu en salle d'urgence il y a
2 jours pour un érythème fessier
et a reçu de la bacitracine
et ___ ___ un antibiotique oral.'

The transformational component at conjunction expansion recovers the full sentences by filling in the gapped information and produces

'Was seen in emergency room
2 days ago for diaper rash
and WAS given bacitracin
and WAS GIVEN oral antibiotic.'

'A été vu en salle d'urgence il y a
2 jours pour un érythème fessier
et a reçu de la bacitracine
et A REÇU un antibiotique oral.'

(where words in capital letters are generated in the course of processing).

And after filling in the empty SUBJECT (shown by empty square brackets), the transformational component produces

'[] Was seen in emergency room
2 days ago for diaper rash
and [] WAS given bacitracin
and [] WAS GIVEN oral antibiotic.'

'[] A été vu en salle d'urgence il y a
2 jours pour un érythème fessier
et [] a reçu de la bacitracine
et [] A REÇU un antibiotique oral.'

The secondary task of the transformational component is to INDEX atoms such as nouns, pronouns, articles and verbs, as well as NSTG and TPOS phrases (stored in node attribute INDEX) for antecedent recovery, and to record TENSE (in node attribute TENSE-ATT).

B.4. REGULARIZATION GRAMMAR

From the output of the transformation component, the regularization component first turns phrases under connectives into Polish notation format (where each pair of square brackets signifies an ASSERTION or FRAGMENT):

```
'and for [was seen in emergency room
          2 days ago]
      [diaper rash]
  and [WAS given bacitracin]
      [WAS GIVEN oral antibiotic].'

'et pour [a été vu en salle d'urgence
          il y a 2 jours]
      [un érythème fessier]
   et [a reçu de la bacitracine]
      [A REÇU un antibiotique oral].'
```

It is assumed here that each transformed ASSERTION or FRAGMENT corresponds to one FORMAT type. At each ASSERTION or FRAGMENT, the component reviews the elements and decides which type of FORMATs fits the phrase, then assigns a node attribute FORMAT-ATT to the ASSERTION or FRAGMENT whose value is the name of the format type decided upon. Each format type has one or more nodes that are characteristic. This helps to formulate a LIST FORMAT-TYPE which will lead us to devise a procedure to pick out in advance which format type an ASSERTION or a FRAGMENT belongs to. This process requires an identification of semantic host and modifiers. In *'episode of fever'*, *'fever'* is the semantic host of the phrase, though *'episode'* is the syntactic host.

B.5. FORMAT GRAMMAR

From a regularized parse tree, the format component creates a format tree corresponding to every ASSERTION or FRAGMENT of the regularized parse tree. The LSP System currently defines three types of format trees: FORMAT1-3 for treatment, FORMAT4 for laboratory tests and results and FORMAT5 for patient description as a result of physical examination and history.

The Format component produces two types of output: a short form of the format tree (where unfilled nodes of the format tree are ignored) and an intermediate form for a standard dBMS called the CTEMPLATE. In the latter form, the results of mapping texts can be displayed in a combined table (CTABLE). The CTABLE for a French *Lettre de Sortie* can be seen in [8].

The format output for the English sentence is shown in Figure 5 and the output for the French sentence in Figure 6. The CTABLE for these sentences is shown in Figure 7. Notice that in the information representation, the corresponding rows of each sentence are almost identical. An exception is the treatment of *'was given'* vs. *'a reçu'*. The English dictionary classed *'give'* as a general medical management verb (TTGEN); thus, it appears in the TXTT column. On the other hand, the translated French verb *'a reçu'* (*'reçevoir'* in the infinitive form) was given no medical subclass; thus it appears in the VERB column. But the crucial information (MED) in the medical treatment column (TXTT) are retained.

Figure 5
Information format for English example

```
* CP_01   1B.01.08
* WAS SEEN IN EMERGENCY ROOM 2 DAYS AGO FOR DIAPER RASH AND GIVEN
* BACITRACIN AND ORAL ANTIBIOTIC .

(CONNECTIVE   (CONJOINED  (CTEXT = 'AND ')))

(CONNECTIVE   (RELATION   (CTEXT = 'FOR ')))

(FORMAT1-3  (TREATMENT   (GEN   (CTEXT = 'WAS SEEN ')
                                (RTEXT = 'IN EMERGENCY_ROOM ')
                                (EVENT-TIME  (TPREP1 (CTEXT = '[P] '))
                                             (NUM (CTEXT = '2 '))
                                             (UNIT (CTEXT = 'DAYS '))
                                             (TPREP2 (CTEXT = 'AGO ')))
                                (TENSE (CTEXT = '[PAST] ')))))

(FORMAT5    (PSTATE-DATA   (S-S  (CTEXT = 'DIAPER_RASH '))))

(CONNECTIVE   (CONJOINED  (CTEXT = 'AND ')))

(FORMAT1-3  (TREATMENT   (GEN   (CTEXT = 'WAS GIVEN ')
                                (TENSE (CTEXT = '[PAST] ')))
                         (MED   (CTEXT = 'BACITRACIN '))))

(FORMAT1-3  (TREATMENT   (GEN   (CTEXT = 'WAS GIVEN ')
                                (TENSE (CTEXT = '[PAST] ')))
                         (MED   (CTEXT = 'ANTIBIOTIC ')
                                (BP-MOD (PTPART (CTEXT = 'ORAL '))))))
```

Figure 6
Information format for French example

```
* CP_01   1B. 1. 8
* A E1TE1 VU EN SALLE DE LE / LA URGENCE IL Y A 2 JOURS POUR UN
* E1RYTHE2ME FESSIER ET A REC4U DE LA BACITRACINE ET UN
* ANTIBIOTIQUE ORAL .

(CONNECTIVE   (CONJOINED  (CTEXT = 'ET ')))

(CONNECTIVE   (RELATION   (CTEXT = 'POUR ')))

(FORMAT1-3  (TREATMENT   (GEN   (CTEXT = 'A E1TE1 VU ')
                                (RTEXT = 'EN SALLE_DE_LE_/_LA_URGENCE ')
                                (EVENT-TIME  (TPREP1 (CTEXT = 'IL Y A '))
                                             (NUM (CTEXT = '2 '))
                                             (UNIT (CTEXT = 'JOURS ')))
                                (TENSE (CTEXT = '[PAST] ')))))

(FORMAT5    (PSTATE-DATA   (S-S (CTEXT = 'E1RYTHE2ME ')
                               (LTEXT = 'UN ')
                               (RTEXT = 'FESSIER '))))

(CONNECTIVE   (CONJOINED  (CTEXT = 'ET ')))

(FORMAT1-3  (TREATMENT   (MED   (CTEXT = 'BACITRACINE ')
                                (LTEXT = 'DE LA ')))
            (VERB     (CTEXT = 'A REC4U ')
                      (TENSE (CTEXT = '[PAST] '))))

(FORMAT1-3  (TREATMENT   (MED   (CTEXT = 'ANTIBIOTIQUE ')
                                (LTEXT = 'UN ')
                                (BP-MOD (PTPART (CTEXT = 'ORAL ')))))
            (VERB     (CTEXT = 'A REC4U ')
                      (TENSE (CTEXT = '[PAST] '))))
```

C. CONCLUSION

Among Indo-European languages there are great similarities in grammar, making it relatively easy to modify the original LSP Medical English grammar to operate on French. Secondly, the great similarities among European languages in respect to technical vocabulary and terminology, especially in medicine, makes it possible to use the sublanguage techniques ("information formatting") of the LSP system for other European languages. The French adaptation is well along; work on German has begun.

Figure 7
Database CTABLEs for English and French examples

ENGLISH EXAMPLE
Was seen in emergency room 2 days ago for diaper rash
and given bacitracin and oral antibiotic.

SID	*ROW*	*CONJUNCT*	*TXTT*	*VERB*	*DIAG SS R*	*PR TM*
01B.01.08	R 01		WAS SEEN IN EMERGENCY_ROOM			[PAST] [P] 2 DAYS AGO
01B.01.08	R 02	"FOR "			DIAPER_RASH	
01B.01.08	R 03	"AND "	WAS GIVEN BACITRACIN			[PAST]
01B.01.08	R 04	"AND "	WAS GIVEN ORAL ANTIBIOTIC			[PAST]

FRENCH EXAMPLE
A été vu en salle d'urgence il y a 2 jours pour un érythème fessier
et a reçu de la bacitracine et un antibiotique oral.

SID	*ROW*	*CONJUNCT*	*TXTT*	*VERB*	*DIAG SS R*	*PR TM*
01B.01.08	R 01		A E1TE1 VU EN SALLE_DE_ LE_/_LA_URGENCE			[PAST] IL Y A 2 JOURS
01B.01.08	R 02	"POUR "			UN E1RYTHE2ME FESSIER	
01B.01.08	R 03	"ET "	DE LA BACITRACINE	A REC4U		[PAST]
01B.01.08	R 04	"ET "	UN ANTIBIOTIQUE ORAL	A REC4U		[PAST]

ACKNOWLEGEMENTS

This research was supported in part by Grant 3.973.0.87 from the *Fonds National Suisse de la Recherche Scientifique.*

REFERENCES

[1] Harris, Z., *Language and Information* (Columbia Univ. Press, New York, 1988).

[2] Kittredge, R., and Lehrberger, J., eds. *Sublanguage: Studies of Language in Restricted Semantic Domains* (Walter de Gruyter, Berlin, 1982).

[3] Proceedings of the Symposium on Computer Applications in Medical Care, (SCAMC). IEEE, New York:
Second Annual (1978) pp. 330-343.
Third Annual (1979) pp. 105-113.
Sixth Annual (1982) pp. 797-804.
Seventh Annual (1983) pp. 688-691, 692-695.
Ninth Annual (1985) 82-86, 221-226.

[4] Sager, N., Friedman, C., Lyman, M.S, Chi, E.C., Macleod, C., Chen, S., and Johnson, S., *The Analysis and Processing of Clinical Narrative*, in: Salamon, R., Blum, B., and Jorgensen, (eds.), MEDINFO 86; Proceedings of the Fifth Conference on Medical Informatics. Elsevier Science Publishers B.V. (North Holland, 1986) pp. 1101-1105.

[5] Sager, N., Friedman, C., Lyman, M.S., and members of the Linguistic String Project, *Medical Language Processing: Computer Management of Narrative Data* (Addison-Wesley, Reading, MA, 1987).

[6] Sager, N. et al, *Adapting a Medical Language Processor from English to French*, submitted to MEDINFO 89.

[7] Borst, F. et al., *Cost Containment and Quality of Care Assessment: By-Product of a Fully Integrated HIS Handling Free Text Analysis of Discharge Summaries*, submitted to MEDINFO 89.

[8] Lyman, M. et al., *Medical Language Processing for Knowledge Representation and Retrievals*, in Proceedings of the 13th Symposium on Computer Applications in Medical Care (1989).

[9] Sager, N., *Natural Language Processing: A Computer Grammar of English and its Applications* (Addison-Wesley, Reading, MA, 1981).