GENERATING MEDICAL CASE REPORTS WITH THE LINGUISTIC STRING PARSER*

Ping-Yang Li, Martha Evens, and Daniel Hier

Department of Computer and
Information Sciences
University of Alabama at Birmingham
Birmingham, AL 35294 (205)934-2213

Computer Science Department
Illinois Institute of Technology
Chicago, IL 60616

Department of Neurology
Michael Reese Hospital
Chicago, IL 60616

ABSTRACT

We are building a text generation module for a decision support system designed to assist physicians in the management of stroke. This module produces multi-paragraph reports on stroke cases stored in the Stroke Data Base or on cases being processed by the decision support system. Analysis of human-generated case reports using Sager's Linguistic String Parser (LSP) led to a characterization of the stroke sublanguage in terms of four components: a Text Grammar for stroke case reports, a set of Stroke Information Formats, a Relational Lexicon for the stroke sublanguage, and a Linguistic String Grammar for this sublanguage. At this point, we have produced free text by using reverse transformations from our LSP grammar to combine fragments into sentences. Our future goal lies in discovering how to generate good paragraphs, using these components as tools.

I INTRODUCTION

Our exhaustive study of stroke case reports has revealed essential information about the stroke sublanguage. Based on this study we have written a Linguistic String Grammar for the stroke sublanguage and a stroke lexicon containing about 3560 entries. We have also developed a set of eleven stroke information formats, which describe the conceptual structures that turn up repeatedly in our reports. As sentences are analyzed with the Linguistic String Parser, they are broken down into elementary assertions, which are then stored in these formats. Inverse transformations from the same grammar are used to combine simple sentences into complex ones in the generation process. In addition we have developed a text grammar which accounts for many of the salient facts about the structure of case reports and which serves as the basis for guiding the process of text generation. In the following paragraphs, we will briefly describe the stroke sublanguage in terms of four components: a Text Grammar for stroke case reports, a set of Stroke Information Formats, a Relational Lexicon for the stroke sublanguage, and a Linguistic String Grammar for this sublanguage.

II THE STROKE LEXICON

The best and most direct way to gain information about the stroke sublanguage is to analyze handwritten case reports generated by physicians. This analysis is the basis of not only the lexicon and the grammar of the stroke sublanguage, but also the semantic classes and the discourse structures, as well as the relations between those classes and attributes.

Our strategy was: first, to generate vocabulary lists and KWIC (Key Word In Context) indices for the texts; secondly, to study KWIC indices to find which words are associated with each other, and to identify lexical-semantic relationships between words. Further steps are then taken to generate word-relation-word triples incorporating the results of the previous steps to record lexical and semantic relationships between words.

One main object of this analysis is to create a Relational Lexicon [Ahlswede and Evens, 1983] containing all the words that might be used in a stroke report. The Relational Lexicon for the Stroke Sublanguage contains both information about words and information about that part of the world we are trying to describe, mainly the anatomy and physiology of the brain. The lexicon is structured as a large network of words connected by arcs representing the relations between them, such as synonymy, taxonomy, part-whole, or relative spatial orientations.

The most familiar example of a lexical-semantic relation is probably synonymy as in

anosognia SYN denial

Some equally important though less familiar relations are taxonomy, the "is-a-kind-of" relation, as in

carotid TAX artery

meaning that the carotid "is a kind of" artery, and the part-whole relation, as in

ventricle PART heart

signifying that a ventricle is a part of the heart. Many other relations appear in less explicit form in ordinary English. The relations LEFT, RIGHT, ABOVE, BELOW, IN-FRONT-OF, and IN-BACK-OF are useful in describing anatomy. The CAUSE relation is particularly useful in explaining reasoning from anatomy to symptoms. We have devised a Relational Lexicon for this sublanguage to record lexical and semantic relationships between words, in the hope of increasing the cohesion of the generated text.

Lesion of left occipital lobe CAUSE alexia

* This work has been supported in part by the AMOCO Foundation through a grant to the Neurology Department at Michael Reese Hospital.

We have also included a number of fixed phrases that occur repeatedly in the stroke case reports such as "visual field", "CT scan", "nursing home", "left sided weakness" and "right sided weakness." Frequently these are phrases which are most conveniently treated as if they were single words. The notion of a phrasal lexicon was suggested by Becker, who proposes that people generate utterances "mostly by stitching together swatches of text that they have heard before." [Becker, 1975, p63] Since the goal of this research is not only to analyze medical case reports but to generate them automatically, the notion of the phrasal lexicon is adopted to facilitate both the parsing and the generation processes. The stroke lexicon, thus, contains not just single words but a number of multi-word phrases that physicians seem to manipulate as a single unit.

III THE LSP GRAMMAR FOR THE STROKE SUBLANGUAGE

The next step in analyzing the stroke reports was to parse them using the LSP. This step served several valuable purposes. First, there is a very close relationship between "parsing grammars" and "generation grammars"; the development of a parsing grammar thus taught us much of what we needed to know to develop the generation grammar. Secondly, the parsed reports exhibit information about the syntactic and even semantic contexts of words and phrases in a systematic way, that is not easy to achieve with unparsed text. Consequently, this greatly simplified the production of the relational lexicon and the text grammar.

The medical sublanguage, we found, deviates from standard English in a number of ways. Neurologists use a number of terms which are not current in ordinary language, and others which are current but which have special meanings in medical contexts. It is full of incomplete sentences. Often the subject is omitted, generally when it is understood to be the patient. Abbreviations are frequent. One typical report begins, "This 47 YO BF was admitted 25 August, 1983 for right sided weakness." "YO" is short for "year old" and "BF" for "black female." As in this example, prepositions are frequently omitted. So are more major parts of speech, particularly verbs, as in, "Lnoxin the only medication." Much of the time the text becomes merely a string of noun phrases. "No CT scan." "No medical rx, as was intolerant to ASA."

A Linguistic String Grammar for the stroke sublanguage has been developed after studying a number of human-generated stroke case reports. The grammar is initially based on Sager's intermediate grammar. By iterative revision, this grammar has been adjusted for the stroke medical texts. Approximately 64 subclasses of the major word classes are currently recognized in the grammar. Special word classes for categories like medications and operative procedures facilitate parsing with our Linguistic String Grammar for stroke reports and help to ensure that sentences are semantically as well as syntactically well-formed. Recently, we have converted this parsing grammar into a generation grammar in order to achieve our goal of generating medical case reports automatically, using techniques suggested by Grishman [1979].

IV STROKE INFORMATION FORMATS

We started with actual case reports and followed the techniques described by Friedman [1983] in developing our information formats for stroke. Each sentence of a case report is eventually categorized into a number of elementary assertions or fragmentary assertions, called information formats. Eventually, we have identified 11 information formats for stroke reports as shown below:

Format 0:	Identification Data
Format 1:	Admission Data
Format 2:	Chief Complaint
Format 3:	Onset of Deficit
Format 4:	Past Medical History
Format 5:	Physical Examination Results
Format 6:	Test Results
Format 7:	Final Diagnosis
Format 8:	Treatment with Drugs
Format 9:	Treatment with Operative Procedure
Format 10:	Discharge Information

Because of the way in which the formats are constructed, there is a close correspondence between word class membership and format column. Once the information formats are constructed on the basis of an analysis of a sample set of case reports, subsequent documents of the same type can be automatically mapped into them. We thus obtain a structured form of the information that is suitable for computerized data processing. Each column or field in an information format contains words or phrases that carry the same kind of information in the texts; each format has certain fixed fields and some of the fields have subfields. Often, formats are connected to each other by conjunctions, prepositions, or other relations. For example, many examples begin with a statement about the patient's admission to the hospital, e.g., "Patient 137 is a 47 year old right-handed black woman admitted for a stroke with mild left-sided weakness." It seems to us that this sentence is a combination of Format-0, Format-1, and Format-2. Since information formats can be considered as a kind of semantic representation for specialized information, this inspired us to use information formats as a base from which to generate sentences.

V THE TEXT GRAMMAR FOR CASE REPORTS

A careful analysis of case reports will not only disclose the grammar of this sublanguage, but also reveal essential knowledge of the text structure. Although our collection of stroke case reports show some wide variations in syntax, the topics reported and the order in which they appear are highly constrained. This has enabled us to devise a text grammar which fits most of our reports very closely and represents the discourse structure of the case reports.

Just as information formats can be thought of as expressing sentence structures of case reports, so can text grammar be thought of as the internal discourse structure of case reports. It is clear that simple sentences are not the highest level of structured linguistic input. Sentences themselves can serve as arguments for higher level organization. In this module, we have developed a text grammar which accounts for many of the salient facts about the structure of case reports and which serves as the basis for guiding the process of text generation. The primary functions of the text grammar are to select information formats and to organize the text content according to the information format. Thus, it will produce an ordered list of the information formats to specify what information to be talked about first, what next, and so forth in an appropriate way.

Figure 1 shows the text grammar that we have developed for our stroke case reports. The paragraph level organization is almost fixed. The first paragraph identifies the patient and describes the chief complaint and the evolution of the deficits. The second paragraph gives relevant information about the patient's past medical history. The next two paragraphs then report the physical examination, and detail the tests performed. Paragraph five shows the result of the final clinical diagnosis which includes the category of the disease, the areas and vessels involved, and the underlying mechanism. If there is more than one diagnosis derived by the decision support system, all alternatives will be listed. The last paragraph states the hospital medication received, and the final outcome which includes the patient's discharge or autopsy information. Although preset paragraph boundaries are embedded in the formulas, they can be dynamically modified depending on the presence of certain symptoms. If the patient, for example, has gone through several operative procedures in the hospital, these will be grouped together in an additional separate paragraph. In this way, a comparatively smooth text can be generated.

VI GENERATION WITH THE LSP

The techniques used in generating free text are based on the Linguistic String Parser [Sager, 1981]. The LSP grammar has two principal components: a BNF grammar and a set of restrictions. The context-free grammar associates with each input sentence a set of parse trees. Restrictions have many functions; one is to state conditions on a parse tree that must be met in order for the tree to be accepted as a correct analysis of the input sentence. These restrictions are used to express detailed wellformedness constraints that are not conveniently statable in the context-free component. In addition, the restriction component contains a number of transformations that decompose a complex sentence into two or more simpler sentences. For instance, the sentence "An echocardiogram showed atrial myxoma and mitral valve lesion." is decomposed into "An echocardiogram showed atrial myxoma." and "An echocardiogram showed mitral valve lesion."

We have taken our Linguistic String Grammar for the Stroke Sublanguage and reversed the transformations using the techniques suggested by Grishman [1979]. A major component of our text generation module is a set of reverse transformational rules derived from our LSP grammar for the stroke sublanguage. The reverse transformational rules consist of a set of aggregation rules and a set of syntactic, semantic, and rhetorical constraints. Both sets of rules function in cooperation to add or delete words from a sentence, reorder the words of a sentence, or combine two sentences to form a larger sentence. We use both simple transformations, which will convert a sentence from one form to another, and complex transformations, which will combine two sentences to form a third. Deletion, substitution, and adjunction are simple transformations which can be thought of as single-sentence transformations. Embedding and conjoining are complex transformations which combine sentences. They can be recursively applied to generate even more complex sentences. The function of an embedding transformation is to take material from a subordinate clause and make it part of the main clause. Conjoining transformations link two coordinate sentences by using conjunctions. In the example below, two sentences, S1 and S2, are merged by using an embedding transformation.

S1: THE PATIENT IS A WOMAN.
S2: THE PATIENT IS BLACK.

```
[Relative Clause Transformation]
$T-RANSFM-1 =
        IF $1 THEN ALL OF $P1, $P2, $P3.

  $1= IF VALUE X1 OF SUBJECT OF
        ASSERTION OF X9 IS NOT EMPTY
       THEN VALUE OF SUBJECT OF
        ASSERTION OF X5 IS X1.

  $P1 = EITHER
        IF X1 HAS ATTRIBUTE NHUMAN
        THEN REPLACE X4 BY SUBJECT X4
        OF ASSERTION OF X9 ('WHO')
        OR
        IF X1 HAS ATTRIBUTE NONHUMAN
        THEN REPLACE X4 BY SUBJECT X4
        OF ASSERTION OF X9 ('WHICH').
  $P2 = REPLACE X7 BY RN X7 OF LNR OF
        NSTG OF SUBJECT OF ASSERTION
        OF X5 (ASSERTION OF X9).
  $P3 = BOTH DELETE X9
        AND  $T-RANSFM-3.
```

This is a simplified transformational rule for relative clauses. To perform this transformation, the system first checks whether the subjects in both sentences are identical. Since this is a global transformational rule, Registers X5 and X9 are used to stand for these two sentences. Once the requirements are satisfied, three operations, $P1, $P2, and $P3, are performed in sequence. Starting with $P1, the system further checks the attributes of this identical subject. If an attribute "NHUMAN", which means the subject is a human being, is found, the system then replaces the subject of Sentence X9 by a relative pronoun, WHO; Otherwise, if a attribute "NONHUMAN" is found, a relative pronoun, WHICH, is introduced. Therefore, we can have many different sentences being generated by this rule, "THE PATIENT WHO ....", and "DIABETES WHICH ...." In $P2, the system copies the modified tree structure of X9 and adjoins it immediately after the subject of X5. Finally, the structure tree of X9 is deleted from the original place and we have the sentence "THE PATIENT WHO IS BLACK IS A WOMAN". The transformational rule, $T-TRANSFM-3, mentioned in $P3 will further transform "THE PATIENT WHO IS BLACK IS A WOMAN" to "THE PATIENT IS A BLACK WOMAN". Further details can be found in [Li et al., 1985].

VII OUR TEXT GENERATION MODULE

The system consists of four components: the Text Structure module, the Information Format module, the Transformation module, and the LSP module. Data from the database is transformed as it flows from one module to the next. In order to manage these components, we have also developed a top-level driver. The top-level driver contains the control information that determines the order in which the components are activated. This monitor also serves as a simple user interface, displaying messages and asking for commands. The text grammar of the stroke sublanguage has been implemented and merged in the Text Structure module which can produce an ordered list of information formats to organize the text content. The Information Format module contains the 11 information formats and the Information Extraction unit. The main tasks of the Information Extraction unit are to infer the numeric data from the database and to map these data into a simple sentence fragment or a series of sentence fragments. Within each information format, there is a set of embedded ordering rules which can organize the information at the sentential level. The Transformation module contains a set of reverse transformational rules. These rules are used to compose a sentence by integrating information from several information formats. The choice of the appropriate transformations is based on the types of sentence fragments available. The LSP module contains Sager's Linguistic String Parser. The following simplified example may be helpful. Initially the system displays welcome messages and asks for entering the report number; that is, the patient's registration number. Control is then passed to the Text Structure module. The topic of the first paragraph is "Init_Info" (Initial_Information) which consists of "Pt_Info" (Patient_Information) and "Dfct_Info" (Deficit_Information). The Text Structure module then first produces and passes an ordered listed of information formats for Patient_Information to the Information Format module. Upon receiving this list, the specified formats are activated and the Information Extraction unit then extracts the desired information from the database and maps it into the appropriate format slots. In Figure 2, Format-0 contains the patient's identification information which includes the patient's registration number, age, handedness (right or left), race (white, black, or oriental), and sex. The information existing in each slot can initially be expressed by a simple primary sentence.

Format 0: Identification Data

Patient	Reg-No	Age	Handness	Race	Sex
PATIENT	423	47	RIGHT	BLACK	FEMALE

S1. THE PATIENT'S NUMBER IS 423.
S2. THE PATIENT IS 47 YEARS OLD.
S3. THE PATIENT IS RIGHT-HANDED.
S4. THE PATIENT IS BLACK.
S5. THE PATIENT IS A WOMAN.

Figure 2. Format-0 and Simple Sentences

These sentences are then parsed by the LSP one at a time. The embedding rule of this format specifies EMBEDDING(APPOSITION(S1), EMBEDDING(S2, EMBEDDING(S3, EMBEDDING(S4, S5)))) as the suggested transformation order. Therefore, the relative clause transformation and apposition transformation are recursely performed by of the Transformation module. We finally obtain the complex sentence "THE PATIENT 423 IS A 47 YEAR OLD RIGHT-HANDED BLACK WOMAN." The linguistic string analysis presented here thus gives us a method of constructing well-formed sentences from certain sentence fragments. Figure 3 shows an example generated by our system.

Clearly we have only begun to explore the possibilities of reverse transformations. Sager's Restriction Language [1981] makes it easy to write and experiment with other transformations. Further study of complex objects, adverbs, and conjunctions will reveal methods of generating a richer set of sentence level structures.

VIII CONCLUSIONS AND FUTURE GOALS

We have produced text by reversing Linguistic String transformations, but our real interest lies in discovering how to generate good paragraphs, using the Text Grammar, the Stroke Information Formats, the Relational Lexicon, and the Linguistic String Grammar as tools.

In the realm of paragraph organization, we are particularly interested in two strategies. Mann's [1981] Fragment-and-Compose paradigm, with its emphasis on building a paragraph from very small linguistic components, is appropriate to our plan of generating text from fragmentary information in information formats and also to the structure of the Relational Lexicon. The other important consideration in generating case reports is deciding what information is to be included in the report and what is to be left out. The salience principle discovered by Conklin and McDonald [1982] in their analysis and synthesis of house descriptions seems to operate as well in medical reports. Within any area the grossest pathology is described first; presumably this is the most salient point from the physician's point of view. Then comes a discussion describing which associated areas are affected.

We want to experiment with more creative ways to use the lexicon. Becker's theory of the phrasal lexicon [1975] tells us that we should be combining long phrases not just individual words. We hope to improve the cohesion of our paragraphs by using the lexical relationships in our Relational Lexicon.

Even our brief experiments with Mandarin and English case reports [Li and Evens, 1985] have suggested that focus mechanisms work differently in these two languages. We want to experiment with the focusing techniques of McKeown's [1982] work in both languages.

Three aspects of our work seem to have particular theoretical interest: the relational lexicon as a knowledge representation structure, the possibilities of the Linguistic String Parser in text generation, and the little-understood problem of text generation at the paragraph level.

REFERENCES

[1] Ahlswede, T. and Evens, M. 1983. "Generating a Relational Lexicon from a Machine-Readable Dictionary," Workshop on Machine-Readable Dictionaries, SRI, April.

[2] Becker, J. 1975. "The Phrasal Lexicon," in Theoretical Issues in Natural Language Processing, eds. R. Schank and B. Nash-Webber, Cambridge, June, 60-63.

[3] Conklin, E.J., and McDonald, D.D. 1982. "Salience: the Key to the Selection Problem in Natural Language Generation," Proc. 20th Annual Meeting of the Association for Computational Linguistics, 129-135.

[4] Friedman, C., Sager, N., Chi, E., Marsh, E., Christenson, C., and Lyman, M. 1983. "Computer Structuring of Free-Text Patient Data," Proc. Seventh Annual Symposium on Computer Applications in Medical Care, IEEE, Washington, D.C., October 23-26, 688-691.

[5] Grishman, R. 1979. "Response Generation in Question Answering Systems," Proc. 17th Annual Meeting of the Association for Computational Linguistics, 99-101.

[6] Li, P.Y., Ahlswede, T., Evens, M., Curt, C., and Hier, D. 1985. "A Text Generation Module for a Decision Support System for Stroke," Proc. of the Conference on Intelligent Systems and Machines, Oakland University, Rochester, MI, April.

[7] Mann, W. and Moore, J. 1981. "Computer Generation of Multiparagraph English Text," American Journal of Computational Linguistics, Vol. 7, No. 2, 17-29.

[8] McKeown, Kathleen R. 1982. Generating Natural Language Text in Response to Questions about Database Structure. Ph.D. Dissertation, U. Penn.

[9] Sager, N. 1981. Natural Language Information Processing: A Computer Grammar of English, Addison-Wesley, Reading, MA.

```
Case_Report %%= Init_Info + Md_Hstry + Phy_Exam +
                 Lab_Tst + Fin_Dex + Outcome
Init_Info   %%= Pt_Info +  Dfct_Evoltn
Pt_Info     %%= Reg_No + Age + Hndnes + Race +
                Sex + Admson + Chf_Complnt
Admson      %%= Disease + (Admson_Dat | Null)
Dfct_Evoltn %%= Onset_Activity  + Dfct_Prgrs +
                Dfct_Symptoms
Dfct_Symptoms %%= Headache + Cnscius_Impair +
                  Vomit + Seizure
Md_Hstry    %%= Hstry_Stroke + Hstry_TIA +
                Hstry_Cardiac + Othr_Arhythm +
                Othr_Md_Hstry
Hstry_Stroke %%= No_Strke + Typ_Strke + Year_Strke
Hstry_TIA    %%= No_TIA + Typ_TIA + TIA_Territory
Hstry_Cardiac %%= Cardiomegaly + Heart_Disease +
                  Atrial_Fbrlatn + Valvular_Lesion
Othr_Md_Hstry %%= Hypcrtension + Diabetes + Coagulopathy
                  + Systemic_Emboli + Arteriosclerosis
Phy_Exam %%= General_Exm + Hghr_Cortcl_Exm +
             Cranial_Exm + Motor_Exm +
             Sensory_Exm + Cerebellar_Exm
Lab_Tst %%= Echocardiogram + Lumbar_Puncture +
            Angiography + Brain_Scan_Flw_Stdy +
            Brain_Scan_Statc_Stdy + Cerebral_Blood_Flw +
            E.E.G. + Complication +  Phonoangiography +
            Oculoplesthymography + Doppler_stdy +
            Cholesterol_Lvl + CT_Scan
Fin_Dex %%= Strke_Category + Vessel_Involved +
            Area_Involved + Mechanism
Outcome %%= Medication + Dschge_Plan
Medication %%= (Med_Drug | Null) + (Med_Surgical | Null)
```

Figure 1. The Text Grammar for Stroke Case Reports

Michael Reese Hospital Stroke Service Report

Patient 165 is a 39 year-old right handed white woman admitted for a stroke with a moderate headache. The deficit came on when she got up in the middle of the night. It was maximal at onset. At the onset of the deficit, there was a moderate headache, a gradual onset of obtundation, and vomiting within the first 12 hours, but no seizure activity.

Past medical history revealed no stroke, TIA, or cardiac disease. There was no evidence that she had systemic emboli or arteriosclerosis.

Examination revealed a lethargic woman with blood pressure of 105/70. There was stiff neck but no carotid bruit. Mental status is normal. Cranial nerve testing showed right ptosis, right Horner's syndrome, and 3rd nerve palsy of right side.

Lumbar puncture showed that CSF was bloody, a CSF xanthochromia 3/10, and a CSF protein 255. The E.E.G. was normal in area appropriate to neurologic deficit. An angiogram of both carotids showed a saccular aneurysm of the right posterior communicating artery. There were no complications of the angiography. A CT scan showed the right ventricular space, and meaningocerebral hemorrhage into the right temporal lobe.

The final clinical diagnosis was subarachnoid hemorrhage. Another possibility was cerebral infarction. The most likely area involved by stroke was the right subarachnoid space. Another possibility was the right temporal lobe. The most likely vessel involved in the stroke was the right posterior communicating artery. The most likely mechanism underlying the stroke was hemorrhage caused by aneurysm.

She died due to stroke but no autopsy was performed.

Figure 3. A Sample Output of the Stroke Case Report Generator