Medical Language Processing with SGML Display

Naomi Sager [1], Ngô Thanh Nhàn [1], Margaret Lyman [2], Leo J. Tick [2]

[1] Courant Institute of Mathematical Sciences, New York University, New York, NY 10012
[2] New York University Medical Center, New York, NY 10016

The paper demonstrates several ways that medical language processing can be combined with emerging display technologies to facilitate the extraction of data from free-text patient documents. The techniques allow rapid review via highlighting of the results of processing. Coupling of text markup with further procedures is envisioned.

INTRODUCTION

Interest and use of the Internet has grown enormously recently. This growth was largely driven by the expansion of the World Wide Web (WWW) as a technique for delivering information. Again much of this was due to the availability of what has come to be called "browsers". These browsers have allowed users relatively easy access to the items available on the WWW. The WWW is now under active study as a methodology for distributing medical information as well as facilitating the care process.

Both the browsers and the production of WWW sites have required the development of sophisticated technologies. In this presentation we describe the use of some of these technologies for the display of medical documents which can be, but may not be, part of a WWW activity.

In the past, one of the barriers to the use of information from patient documents has been the lack of generally available tools for handling and displaying text in any but the most straightforward linear mode. Today, software based on Standard Generalized Markup Language (SGML) and Hypertext Markup Language (HTML), known to many through the WWW or through desk top publishing [1-3], appears to remove this limitation. Some think that SGML technology will change how medical information systems will be designed in the future [4].

As described by Tom Lincoln [5]: "Document processing using the (ISO) Standard Generalized Markup Language (SGML) considers each component of the medical chart as a loosely structured document where the document outline can be uniquely delimited in some uniform manner by tags or labels. SGML has been designed for this purpose with respect to data display and formatting, and there are HL-7 compatible approaches to extend these conventions to organize medical content. Tagging the outline can capture each specific information chunk in a flexible format suitable for subsequent retrieval and processing. Other tags and coordinating mechanisms can then be built up on this basis."

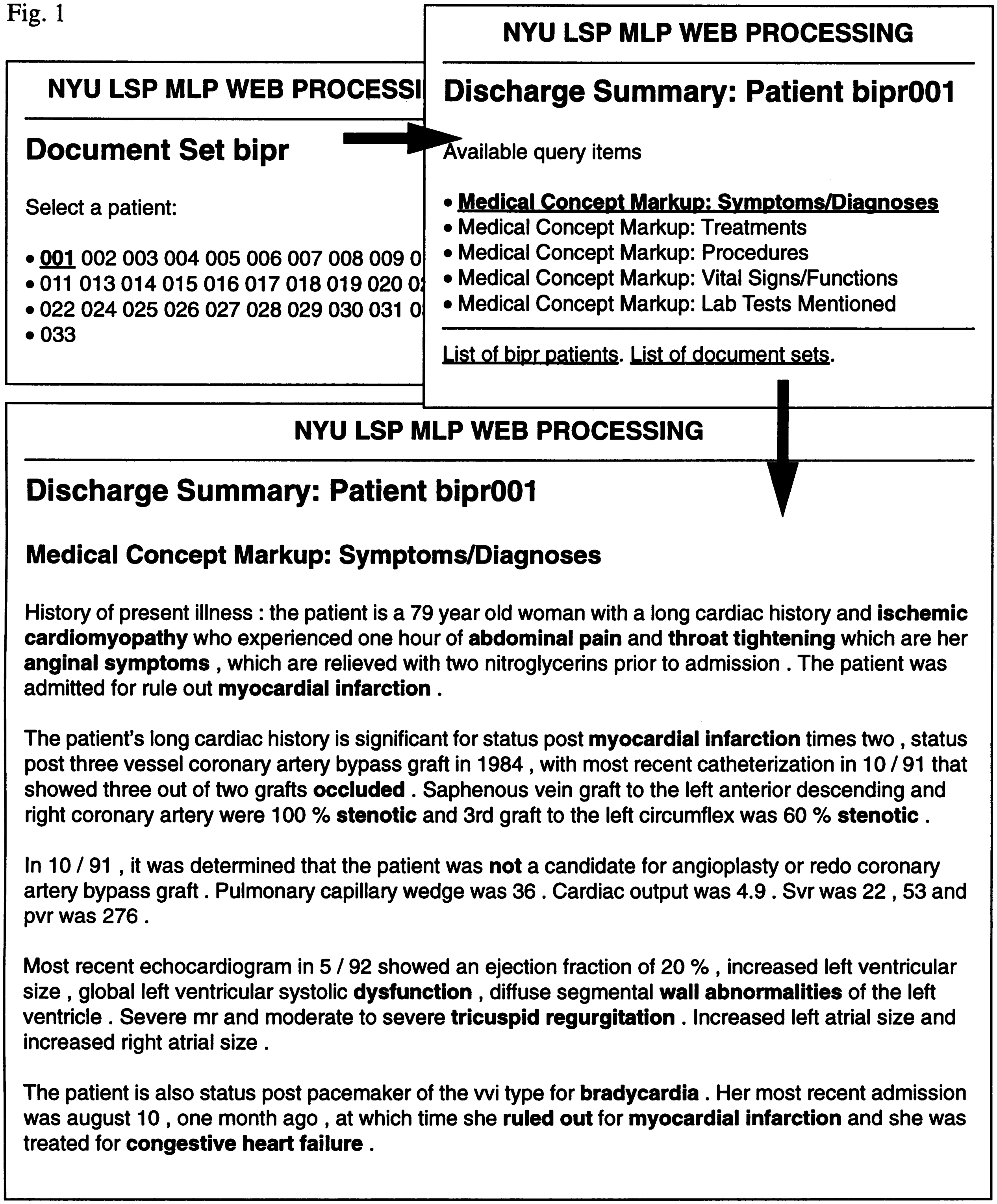
This paper addresses in a preliminary manner some of the ways in which medical linguistic resources and medical language processing can be combined with SGML to facilitate the extraction of patient data for particular applications from routinely collected free-text patient reports. The user could be a manual coder scanning a document for codable items, or a person performing a medical audit on the treatment of a particular problem in many patients, using discharge summaries as a point of departure. The common feature is the better use of human resources - making it easier for people to focus quickly on relevant information when that information occurs in textual form.

LINGUISTIC MARKUP OF MEDICAL TEXT

Narrative documents collected in the course of patient care could supply some of the information required for outcomes research, medical audit, coding, and research databases, provided that it becomes feasible to extract what is relevant for these purposes from large amounts of text. Efforts devoted to medical language processing (MLP) [6-12] and medical vocabulary issues [13-15] are beginning to provide the tools needed for the task.

Two types of linguistic markup are presented here, one simple, the other more complex. The first, Medical Concept Markup, uses only a lexicon and achieves its utility by the relevance of the semantic categories of the lexicon to the types of information requested by the user. The second, Data Extraction Markup, uses the full power of medical language processing,

Fig. 1



NYU LSP MLP WEB PROCESSI

Document Set bipr

Select a patient:

- <u>001</u> 002 003 004 005 006 007 008 009 0
- 011 013 014 015 016 017 018 019 020 0
- 022 024 025 026 027 028 029 030 031 0
- 033

NYU LSP MLP WEB PROCESSING

Discharge Summary: Patient bipr001

Available query items

- **<u>Medical Concept Markup: Symptoms/Diagnoses</u>**
- Medical Concept Markup: Treatments
- Medical Concept Markup: Procedures
- Medical Concept Markup: Vital Signs/Functions
- Medical Concept Markup: Lab Tests Mentioned

<u>List of bipr patients</u>. <u>List of document sets</u>.

NYU LSP MLP WEB PROCESSING

Discharge Summary: Patient bipr001

Medical Concept Markup: Symptoms/Diagnoses

History of present illness : the patient is a 79 year old woman with a long cardiac history and **ischemic cardiomyopathy** who experienced one hour of **abdominal pain** and **throat tightening** which are her **anginal symptoms** , which are relieved with two nitroglycerins prior to admission . The patient was admitted for rule out **myocardial infarction** .

The patient's long cardiac history is significant for status post **myocardial infarction** times two , status post three vessel coronary artery bypass graft in 1984 , with most recent catheterization in 10 / 91 that showed three out of two grafts **occluded** . Saphenous vein graft to the left anterior descending and right coronary artery were 100 % **stenotic** and 3rd graft to the left circumflex was 60 % **stenotic** .

In 10 / 91 , it was determined that the patient was **not** a candidate for angioplasty or redo coronary artery bypass graft . Pulmonary capillary wedge was 36 . Cardiac output was 4.9 . Svr was 22 , 53 and pvr was 276 .

Most recent echocardiogram in 5 / 92 showed an ejection fraction of 20 % , increased left ventricular size , global left ventricular systolic **dysfunction** , diffuse segmental **wall abnormalities** of the left ventricle . Severe mr and moderate to severe **tricuspid regurgitation** . Increased left atrial size and increased right atrial size .

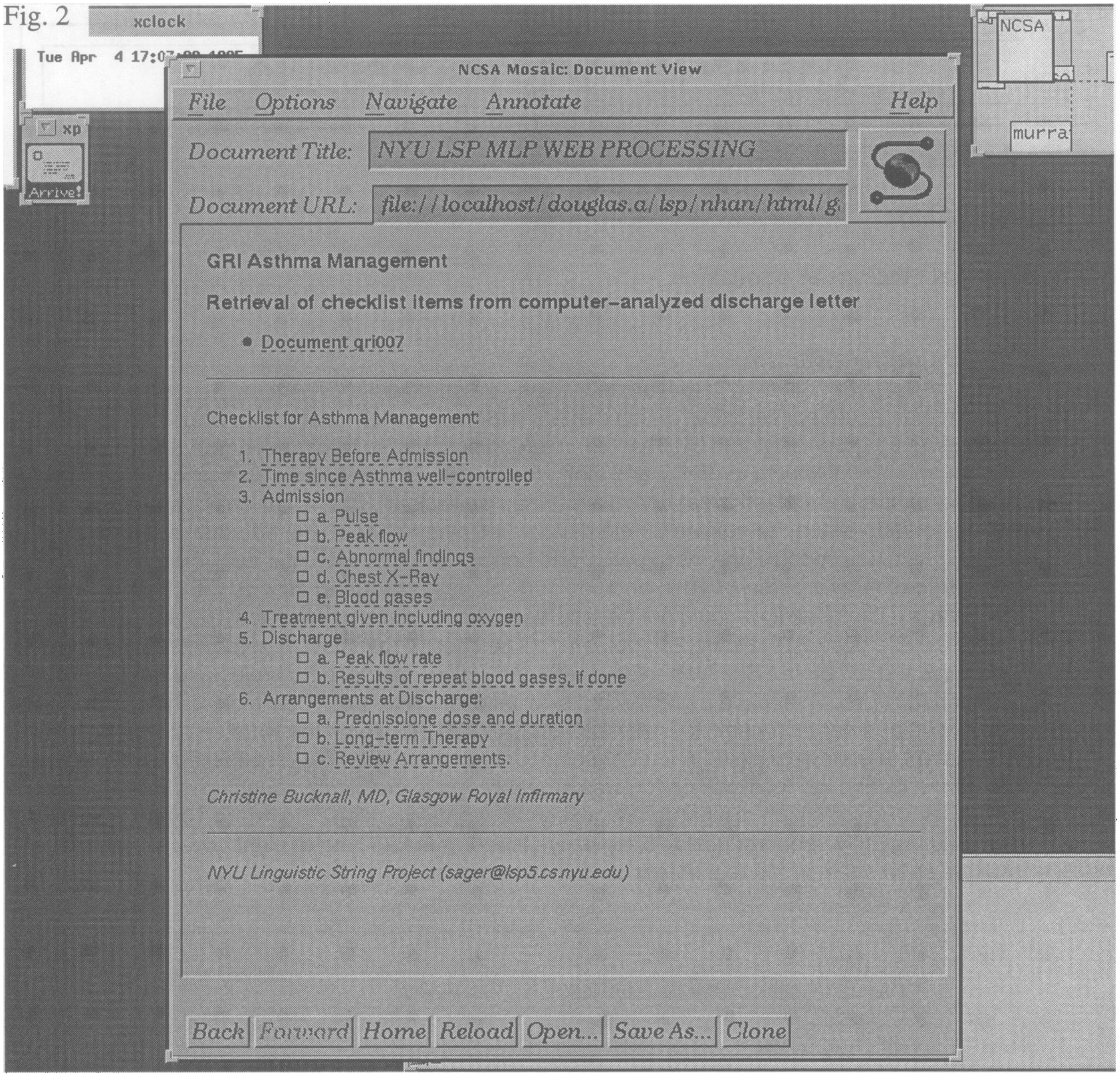
The patient is also status post pacemaker of the vvi type for **bradycardia** . Her most recent admission was august 10 , one month ago , at which time she **ruled out** for **myocardial infarction** and she was treated for **congestive heart failure** .

including parsing, semantic "disambiguation", and the ability to be coupled to complex retrieval procedures.

Medical Concept Markup

The Institute for the Computer-based Patient Record (CPRI) conducted 2 exercises in 1993 in which a group of experts extracted what they deemed to be the important medical concepts in 2 sets of clinical documents [17]. These medical concepts were found to correspond closely to medical semantic classes in the lexicon developed by the New York University Lin-

Fig. 2



GRI Asthma Management

Retrieval of checklist items from computer-analyzed discharge letter

- Document gri007

Checklist for Asthma Management:

1. Therapy Before Admission
2. Time since Asthma well-controlled
3. Admission
   - □ a. Pulse
   - □ b. Peak flow
   - □ c. Abnormal findings
   - □ d. Chest X-Ray
   - □ e. Blood gases
4. Treatment given including oxygen
5. Discharge
   - □ a. Peak flow rate
   - □ b. Results of repeat blood gases, if done
6. Arrangements at Discharge:
   - □ a. Prednisolone dose and duration
   - □ b. Long-term Therapy
   - □ c. Review Arrangements.

Christine Bucknall, MD, Glasgow Royal Infirmary

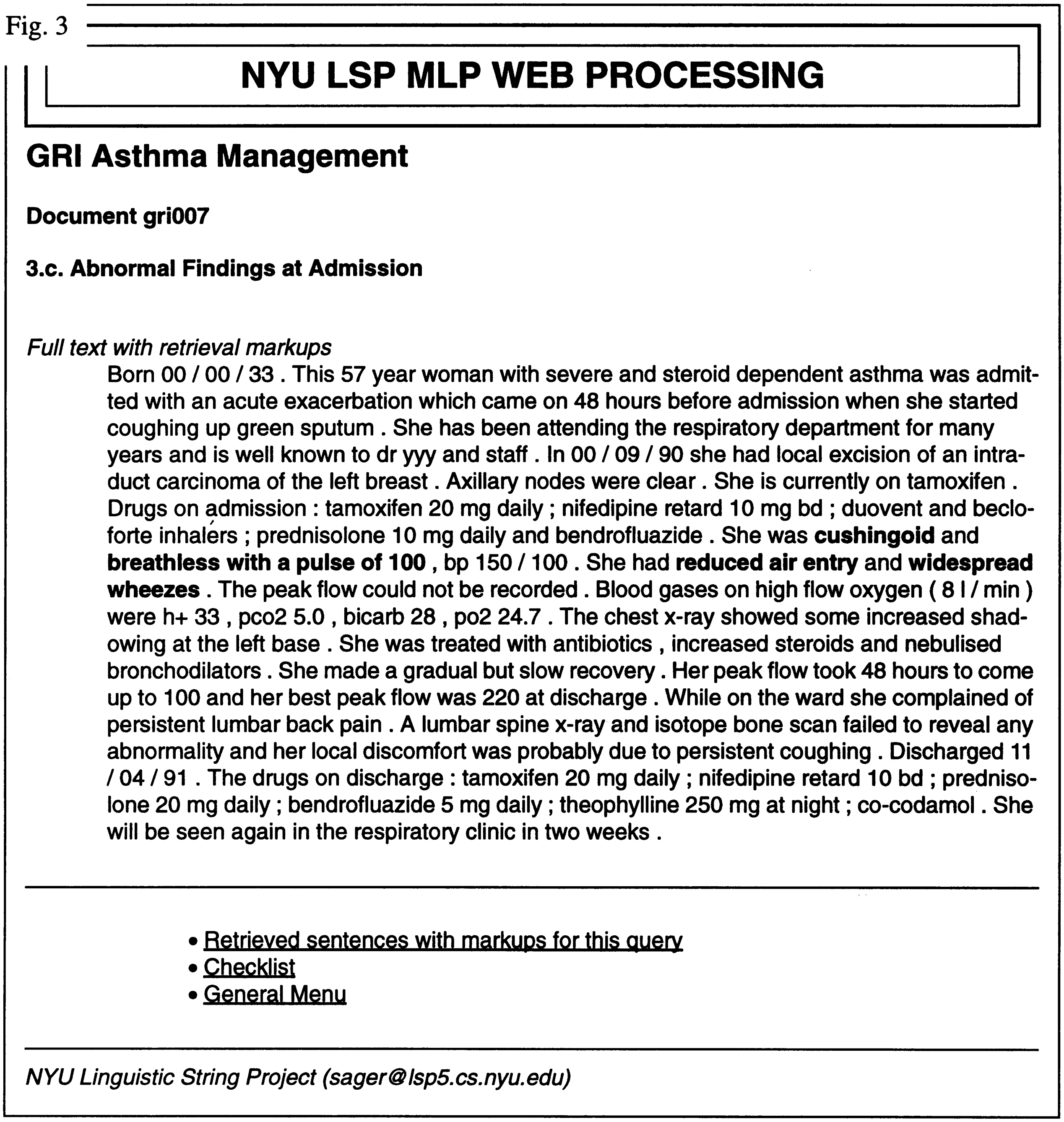
NYU Linguistic String Project (sager@lsp5.cs.nyu.edu)

guistic String Project (LSP) for use in its MLP system [18]. A program was written to embed SGML directives in documents so as to highlight the occurrences of medical concepts. A word or word group was marked if, on lookup in the LSP lexicon, the word or word group was found to have been assigned to a semantic class that had been selected for use in the program that produces the Medical Concept Markup display.

Figure 1 shows successive screens leading to the display of "Medical Concept Markup: Symptoms/Diagnoses" in a discharge summary of a patient with a diagnosis of myocardial infarction. The document was 3 pages long, not uncommon for patients with this diagnosis. The highlighted word classes are the LSP H-INDIC (disease indicator = symptom), H-DIAG (diagnosis), H-PTPART (anatomy) and H-NEG (negation). The anatomy words are not highlighted unless adjacent to a symptom or diagnosis word. Negation words are always highlighted. [Note that the tenseless verb "*rule out*" in paragraph 1 of Fig. 1 is correctly not highlighted as a negation word, whereas "*ruled out*" in the past tense is a negation of a symptom/diagnosis and is seen highlighted in paragraph 5.]

The current LSP English medical lexicon contains ca. 25,000 words coded for their syntactic and medical semantic properties. Outside sources utilized include

Fig. 3



NYU LSP MLP WEB PROCESSING

GRI Asthma Management

Document gri007

3.c. Abnormal Findings at Admission

Full text with retrieval markups

Born 00 / 00 / 33 . This 57 year woman with severe and steroid dependent asthma was admitted with an acute exacerbation which came on 48 hours before admission when she started coughing up green sputum . She has been attending the respiratory department for many years and is well known to dr yyy and staff . In 00 / 09 / 90 she had local excision of an intraduct carcinoma of the left breast . Axillary nodes were clear . She is currently on tamoxifen . Drugs on admission : tamoxifen 20 mg daily ; nifedipine retard 10 mg bd ; duovent and becloforte inhalers ; prednisolone 10 mg daily and bendrofluazide . She was **cushingoid** and **breathless with a pulse of 100** , bp 150 / 100 . She had **reduced air entry** and **widespread wheezes** . The peak flow could not be recorded . Blood gases on high flow oxygen (8 l / min) were h+ 33 , pco2 5.0 , bicarb 28 , po2 24.7 . The chest x-ray showed some increased shadowing at the left base . She was treated with antibiotics , increased steroids and nebulised bronchodilators . She made a gradual but slow recovery . Her peak flow took 48 hours to come up to 100 and her best peak flow was 220 at discharge . While on the ward she complained of persistent lumbar back pain . A lumbar spine x-ray and isotope bone scan failed to reveal any abnormality and her local discomfort was probably due to persistent coughing . Discharged 11 / 04 / 91 . The drugs on discharge : tamoxifen 20 mg daily ; nifedipine retard 10 bd ; prednisolone 20 mg daily ; bendrofluazide 5 mg daily ; theophylline 250 mg at night ; co-codamol . She will be seen again in the respiratory clinic in two weeks .

- Retrieved sentences with markups for this query
- Checklist
- General Menu

NYU Linguistic String Project (sager@lsp5.cs.nyu.edu)

Snomed III.2 and lists of medications, organisms, and abbreviations.

Data Extraction Markup

In Data Extraction Markup, a user's application utilizes full medical language processing (parsing and other linguistic procedures) to extract specific information from free-text documents. The results of the data extraction are then converted to embedded directives in the original documents so that the segments of text that gave rise to extracted data can be reviewed as highlighted segments of text in the original document. In the use of Data Extraction Markup shown here, the application was to determine whether asthma discharge summaries contained documentation of items deemed important for quality assurance of asthma care [11]. The Glasgow Royal Informary supplied the LSP with a set of 59 asthma discharge summaries and a checklist of asthma management items (Fig. 2). The documents were analyzed by the LSP medical language processor (LSP-MLP) and the results were mapped into a relational database. SQL retrieval queries were executed to determine if the quality assurance requirements were reported in the documents.

Words in a row (or rows) of the database that answered a query were then highlighted in the document using SGML markup.

Figure 2 shows the screen which results from the choice of Document gri007. By clicking on one of the checklist items, for example item 3c (Admission Abnormal Findings), the document is displayed with the "answer words" highlighted, as shown in Fig. 3. If no answer is obtained, the document is presented without highlighting and the user can determine if the fault was in the processing or in the original document.

FUTURE ENHANCEMENTS

The benefit of combining linguistic resources with SGML techniques will be better realized when software becomes available to allow operations on the marked text, i.e. hooks to other procedures. For example, particular combinations of symptoms could be the mechanism by which one would retrieve similar cases. As another example, categories like SYMPTOMS/DIAGNOSES and TREATMENTS, used for highlighted displays in the work reported here, could also serve as input to an automatic encoder. Literature search in response to combinations of highlighted items is clearly another potential application.

References

1. Special Section: Hypermedia, *Comm ACM* 37:2, Feb. 1994.
2. Graham IS. *HTML Sourcebook*, Wiley, New York, 1995.
3. Cover story: Electronic Publishing, *PC Magazine* 14:3, Feb.7, 1995, 110-162.
4. Lincoln T, Essin DJ, Anderson R, Ware W. The Introduction of a new Document Processing Paradigm into Health Care Computing, submitted to CAIT Nov 1, 1994 and distributed to Soc. Med. Inf. on the Internet and the ACMI listserv.
5. Communication on the ACMI listserv.
6. Sager N, Friedman C, Lyman MS. *Medical Language Processing: Computer Management of Narrative Data.* Reading, MA: Addison-Wesley, 1987.
7. Lin R, Lenert L, Middleton B, Shiffman S. A free-text processing system to capture physical findings: Canonical Phrase Identification System (CAPIS). Proc Annu Symp Computer Applications in Medical Care. 1991; 15:843-847.
8. Baud RH, Rassinoux A-M, Scherrer J-R. Natural language processing and semantical representation of medical texts. Meth Inform Med. 1992;31:117-125.
9. Satomura Y, Do Amaral MB. Automated diagnostic indexing by natural language processing. Med Inf (Lond) 1992;17:149-163.
10. Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Comput Biomed Res. 1993;26:467-481.
11. Sager N, Lyman M, Bucknall C, Nhàn N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Informatics Assoc. 1994;1:142-160.
12. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Informatics Assoc. 1994;1:161-174.
13. Coté RA, Rothwell DJ, Beckette R, Palotay J, eds. *SNOMED International.* Northfield, IL: College of American Pathologists, 1993.
14. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Informatics Assoc. 1994;1:35-50.
15. McCray AT, Srinivasan S, Browne AC. Lexical Methods for Managing Variation in Biomedical Terminologies, *Proc Annu Symp Computer Applications in Medical Care.* JAMIA Symp Suppl 1994, 235-239.
16. Sager N, Lyman M, Nhàn NT, Tick LJ. Medical Language Processing: Applications to Patient Data Representation and Automatic Encoding. Meth Inform Med. 1995;34:1-2, 140-146.
17. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR for the Computer-Based Patient Record Institute's Work Group on Codes & Structures. The content coverage of clinical classifications. J Am Med Informatics Assoc. 1996; 3:224-233.
18. Sager N, Lyman M, Nhàn NT, Tick LJ. Automatic Encoding into SNOMED III: a Preliminary Investigation, *Proc Annu Symp Computer Applications in Medical Care.* JAMIA Symp Suppl 1994, 230-234.