Using phrases generated by a shallow syntactic parser to match UMLS Metathesaurus concepts in a radiology test collection

by

Mark Francis Mailhot, M.D.

A Thesis

Presented to the Department of Medical Informatics
and the Oregon Health Sciences University School of Medicine
in partial fulfillment of
the requirements for the degree of
Masters of Science
April 24, 2001

TABLE OF CONTENTS

Acknowledgements

I would like to thank the following people, without whom this thesis would not have been accomplished.

My Lord and Savior, Jesus Christ.

My thesis advisor, Dr. William Hersh, and my thesis committee members, Dr. Judith Logan and Dr. Susan Price. Their advice and expertise was invaluable.

My friends, especially Yang Gong.

My family, especially my mother and father.

My roommates and former roommate, Matt, Ryan, and Eric.

Abstract

Introduction

Text forms a significant part of the medical record. Being able to map text to concepts in a controlled medical vocabulary could be useful for teaching, patient care, and research. Natural Language Processing (NLP) systems have figured prominently in attempts to process medical text, particularly radiology dictations. Chart Engine, an existing system for mapping radiology text to Unified Medical Language System (UMLS) Metathesaurus concepts, currently uses barrier words — a "poor man's NLP parser" — to select candidate noun phrases for mapping. In this thesis, Chart Engine was equipped with a syntactic parser as a preliminary processing step that would then supply candidate noun phrases instead of the phrases supplied by the list of barrier words. Several different types of noun phrases were tested.

Methods

We studied a test collection of 50 radiology documents indexed to concepts in the Metathesaurus by a medical librarian. Using EngLite, a shallow syntactic parser, three sets of candidate noun phrases were formed from the documents and then mapped to Metathesaurus concepts using Chart Engine. The results were compared with the baseline data using the same process except that Chart Engine was given noun phrases derived from a list of barrier words. Precision (the number of index concepts found divided by the total number of concepts found) and recall (the number of index concepts found divided by the total number of possible index concepts) were calculated and compared.

Results

Maximal simple noun phrases input to Chart Engine had little effect on either precision or recall. Allowing both maximal simple noun phrases and their subphrases increased recall somewhat but decreased precision. Addition of trailing prepositional phrases had little effect on recall, while again decreasing precision.

Conclusion

Using simple syntactic methods to enhance a text-mapping tool showed some positive results on recall at the expense of precision. An approach that utilized additional techniques would be needed to increase both recall and precision.

Introduction

Background

Text forms a significant part of the medical record. Parts of the record commonly found as free, unstructured text include the admission summary, progress notes, radiology dictations, pathology reports, and discharge summaries. Increasingly, this text is available electronically. (Friedman et al., 1994)

The ability to process this text by computer could be useful for many purposes, including information retrieval, quality assurance, and outcomes research. The ability to process the text of radiology dictations, in particular, could be useful for teaching, patient care, or research. Because radiology dictations serve as stand-ins for the information contained in the X-rays, computed tomography scans (CTs), and other images, if one were able to process this text, one could have ready access to information about the images themselves. (Lowe et al., 1995; Lowe et al., 1998) So, for example, the ability to find chest X-rays where tracheal deviation was noted could be useful to a trauma attending physician teaching residents to recognize signs of a tension pneumothorax. The ability to find Magnetic Resonance Imaging scans (MRIs) in which an adrenal mass was seen, in conjunction with other information linked to the patient record, could be useful to a radiologist trying to evaluate a given adrenal mass. The ability to find all chest X-rays in which a pulmonary nodule was remarked upon could be useful to a public health scientist investigating the factors associated with follow-up of incidental pulmonary nodules.

Beginning with a research system for indexing radiology text, Chart Engine, we modified the system to incorporate Natural Language Processing (NLP) techniques. The original system used a set of barrier words, common English words unlikely to contain any medical meaning, to form phrases for mapping to the Unified Medical Language System (UMLS) Metathesaurus. We hypothesized that using an off-the-shelf English sentence analyzer as opposed to the set of barrier words to form these phrases would be feasible, and that in doing so we could improve the performance of the system.

Background – Natural Language Processing

The technique we chose to use in this work is called syntactic parsing. Syntactic parsing is part of a bigger set of tools that form the field of NLP. Another name for NLP is computational linguistics, which has been defined as "the subfield concerned with computer programs to understand and generate natural language." (Hersh, 1996) NLP has been an area of exploration within the field of computer science for over 30 years.

NLP takes advantage of the fact that natural language (e.g., English) contains a hierarchy of units of meaning. For our purposes, the basic unit is the *lexeme*. A lexeme is a word, abbreviation or acronym. Lexemes come together to form *chunks*. For example, "myocardial infarction following" is a chunk made up of three lexemes. *Tokens* consist of either lexemes or punctuation.

Certain chunks of lexemes are designated *phrases*. Chunks like “the heart,” “the solar eclipse,” “the quiet man,” and “the man who knew too much” are examples of *noun phrases* because they can take the place of a noun in a sentence. For example, it is grammatical to say, “Here is the heart” but ungrammatical to say “Here is the myocardial infarction following.” Chunks like “played,” “ate the sandwich,” and “ran madly down the road” are examples of *verb phrases*. Certain chunks headed by a preposition, such as “in the street,” “down by the bay,” or “by the nape of his neck” are known as *prepositional phrases*.

It should be possible to take any proper English sentence and divide it into phrases, and to continue subdividing the phrases until we reach the level of lexemes. In Figure 1, the sentence “The man played with the ball” is first divided into the noun phrase (NP) “the man” and the verb phrase (VP) “played with the ball.” The phrase “played with the ball” is subdivided into the verb (V) “played” and the prepositional phrase (PP) “with the ball.” Note that “played” is a verb, but also a verb phrase — it would be perfectly grammatical to say, “The man played.” Going further, the phrase “with the ball” is subdivided into the preposition (PREP) “with” and the noun phrase “the ball.” The remaining noun phrases, “the man” and “the ball,” can be further subdivided into a determiner (“the”) and a noun (“man” or “ball,” respectively).

Syntactic NLP makes use of rules governing phrases. For instance, one rule for English states that a sentence (S) can be made up of a noun phrase (NP) followed by a verb phrase (VP). One can write this rule down symbolically as follows:

S ← NP VP

Another rule might state that a noun phrase (NP) can consist of a determiner (DET) and a noun (N):

NP ← DET N

or that a verb phrase (VP) can consist of a verb (V):

VP ← V

NLP using *syntactic* techniques generally takes advantage of a body of such rules, along with a *lexicon*. The lexicon gives rules about what lexemes are allowed as nouns, verbs, prepositions and other *parts of speech*. For example, a simple lexicon might state that:

"the" is a determiner
"man" can be a noun or a verb ("to man the boat")
"ate" is a verb.

With this lexicon and the above three rules, we can deduce the fact that "The man ate" is a legitimate sentence.

Deducing in a logical fashion that this sentence is legitimate is known as *parsing*. In general, a computer program or algorithm designed to analyze language from a set of rules is known as a *parser*.

Semantic techniques in NLP take into account not only the parts of speech of lexemes but their meaning as well. For example, a parser equipped with proper semantic knowledge would be able to recognize that "The shoe cried" is a nonsense sentence, even though it is syntactically correct. Thus *semantics* represents a deeper level of meaning in language than syntax. Like syntax, semantic knowledge can be approximated by a set of rules.

At a deeper level than semantics is *pragmatics*, or world knowledge. For example, the sentence "My four year-old son ran a marathon yesterday" is syntactically and semantically correct, but pragmatically highly unlikely. Pragmatics involves knowing that a marathon is a race of over 26 miles, and that four year-olds, as a rule, are not capable of running such a distance.

Within the last 15 years, there has been considerable interest in Medical Language Processing (MLP), the application of Natural Language Processing to the more limited domain of medical text. Because medical language is a subset of language, or *sublanguage*, it is generally acknowledged that processing medical language poses an easier task than processing natural language in general.

Current NLP Systems in Medicine

There have been several notable medical NLP efforts in the English language. One of the earliest efforts to apply NLP to medicine was the Linguistic String Project begun by Naomi Sager. What started out as a general-purpose processor was applied to the field of medicine. (Sager et al., 1994) Published applications include querying asthma discharge reports for quality assurance. (Sager et al., 1993)

One of the most well known medical NLP systems is called MEDLEE. Initially designed to process chest X-rays (Friedman, 1994), MEDLEE has since been expanded to include other types of narratives. (Friedman, 1997) The system maps text to concepts in the Medical Entities Dictionary (MED), a proprietary medical vocabulary. Hripcsak et al. showed that this system, when trained, performed on par with six internists and six radiologists in being able to "diagnose" six clinical conditions, including acute bacterial pneumonia and chronic obstructive pulmonary disease, from chest X-ray dictations. (Hripcsak et al., 1995) In subsequent paper, Hripcsak et al. showed that an event monitor based on analysis of chest X-ray dictations could be useful in improving tuberculosis respiratory isolation compliance. (Hripcsak et al, 1997)

At the LDS Hospital in Utah, the Special Purpose Radiology Understanding System (SPRUS) was developed to interface with the HELP Hospital Information System. Like MEDLEE, SPRUS was originally designed to process chest X-ray dictations, but since was expanded to include other types of text. One application of SPRUS, in concert with the HELP system, has been the automatic encoding of admission diagnoses based on limited admission information. (Gundersen et al., 1996)

An initial version of SPRUS made use of “semantic” information as it parsed chest X-rays. That is, it used the expectation that given findings would occur in given locations or patterns when associated with given diseases. SPRUS looked for certain key words relating to findings or diseases (e.g., “infiltrate”), and then attempted to fill in other expected particulars about the findings or diseases (e.g., “alveolar infiltrate,” “diffuse infiltrate”). (Haug et al., 1990) A later version of SPRUS incorporated information about English grammar. (Gundersen et al., 1996)

Lin et al. (1992) described a system called the Canonical Phrase Identification System (CAPIS). CAPIS analyzed text on a word-by-word basis, using an algorithm to identify phrases from the text and determine whether they were abnormal findings or were normal. (In this way it functioned similarly to the baseline Chart Engine system upon which our work was based). Phrases were then *normalized* using a thesaurus that associated a series of words with similar meaning (e.g., “cancer,” “mesothelioma,” “lymphoma,” “malignancy”) with one “concept word” (“cancer”). As an example, the phrases “carcinoma in the left lung” and “left lung malignancy” might both be normalized to the phrase “left lung cancer.” The normalized phrases were then matched up against a set of *canonical phrases*, pre-defined phrases that had special meaning to the system. The algorithm first looked for exact matches between the normalized phrase and a canonical phrase, then for partial matches. Under certain circumstances other normalized phrases from the same sentence were added to the first normalized phrase until an appropriate match was found. Zingmond and Lenert (1993) used CAPIS to implement a system for monitoring chest X-rays, and showed potential clinical benefit.

Two other recent systems have used different tactics for selecting “candidate phrases.” MetaMap (Aronson, 1996; Tuttle et al., 1998) makes use of syntactic methods, specifically the Specialist Minimal Commitment Parser, which is part of the UMLS distribution. Finally, Nadkarni et al. (2001) describe a system for mapping text to Metathesaurus concepts, which uses IBM’s Feature Extraction Tool (FET), employing a combination of NLP heuristics and a user-defined list of barrier words, to select candidate phrases for mapping.

The systems described above are only a fraction of those that have been developed to extract information from text. They all employ varying degrees of NLP. However, there is a broad spectrum of applications to process text, using a variety of techniques. Referring specifically to the problem of *indexing* text, both MEDLEE and the system described by Nadkarni, et al. map text to a controlled medical vocabulary. However, many systems index text by words alone, or by phrases that either exist in the text or are inferred from the text. Given that one wishes to index text with a controlled medical vocabulary, there are still a number of vocabularies to choose from. The MEDLEE system mapped to the Medical Entities Dictionary, a proprietary vocabulary, while the system described by Nadkarni et al. mapped to the UMLS Metathesaurus. Finally, given the choice of a vocabulary, one must find a *means* to go from text to this vocabulary. It is here that Natural Language Processing techniques can be applied.

Prior Work With Chart Engine

At OHSU, in conjunction with researchers at the University of Pittsburgh, we developed a text processing application called Chart Engine whose purpose it is to index radiology dictations. This is a series of computer programs that can process the text of a radiology dictation and return a list of concepts from the UMLS Metathesaurus (Lindberg et al, 1993) that are appropriate for that report. The goal is to represent the relevant findings, features, and procedures described in the dictation by these concepts. (Lowe et al., 1998)

We had previously trained and tested the system using a collection of 50 radiology dictations, comprising bone scans, chest X-rays, CT scans of the chest, the abdomen and the pelvis, CT scans of the head, and head MRIs. A medical librarian had independently assigned two to 20 Metathesaurus concepts to each scan, each scan receiving 10 concepts on average. These indexing concepts served as a gold standard against which we compared the performance of our system. A sample dictation and its indexing concepts are found in Appendix A.

For development, we had set aside 16 of the dictations as a training set that we were allowed to examine and analyze. The remaining 34 dictations served as a test set. During development, we did not look at the dictations in the test set, nor did we look at the indexing concepts assigned to them. Development of Chart Engine was an iterative process with cycles of failure analysis, hypothesis generation, development and testing. Our primary outcome measures were precision and recall, which we calculated for both the training and test sets of dictations. However, the definitions of precision and recall differed slightly from their traditional definitions in the field of Information Retrieval. We defined

$$\text{Precision} = \frac{\text{the number gold standard concepts successfully retrieved for a given scan}}{\text{the number of unique Metathesaurus concepts retrieved for the scan}}$$

and

$$\text{Recall} = \frac{\text{the number gold standard concepts successfully retrieved for a given scan}}{\text{the number of gold standard concepts assigned to that scan}}$$

Precision or recall for a collection of documents was defined as the weighted average precision or recall, i.e., the sum of the individual numerators used to calculate precision (or recall) divided by the sum of the individual denominators.

Our final algorithm for Chart Engine used a process employing simple linguistic techniques, a tool to map free text to concepts (SAPHIRE) (Hersh and Leone, 1995), information contained in the UMLS Metathesaurus and Semantic Network, and several algorithms and heuristics.

The details of the Chart Engine algorithm were as follows. For each text report, we first “cleaned” the text, getting rid of periods (such as decimal periods) that most likely did not represent breaks between sentences. Some additional punctuation (for example, the slash (“/”) in “s/p” and “h/o”) was removed as well.

Then we grouped the text into sentences. If a sentence contained a statement like “CT of the chest” or “MRI of the brain” (we looked for the words “CT” or “MRI” with or without the word “scan” or “images,” or for the words “computerized” or “magnetic”) we removed the words “of” and “the.” Based on an educated guess about where historical information (“Patient has a history of breast cancer”) occurs in a document, and what words denote a transition to the descriptive part of the document, we also removed from zero to six sentences from further consideration. We did this because in general we were not concerned with indexing concepts that occurred as historical information.

With each remaining sentence, we divided the sentence into a series of phrases, based on the list of 290 barrier words (Appendix B). Another algorithm then sought to determine if phrases were negated (e.g., “There is no hilar adenopathy”) or normal (e.g., “The heart contour is normal”). The phrases that were not negated or normal were considered “candidate phrases.”

We then took the list of candidate phrases and sent them to SAPHIRE. Each phrase was returned with a list of Metathesaurus concepts, each concept having a unique ID number, a text string representing its canonical (preferred) form, and a weight measuring the degree of similarity between the concept and the phrase. Concepts were returned in descending order of weight — higher weights expressing a greater degree of similarity between the concept and the phrase. Thus the better matches were returned at the beginning of the list. Some phrases came back with no concepts; other phrases produced a list of 30 or more concepts.

After getting the list of phrases and concepts back from SAPHIRE, we went through the list and looked at each phrase-concept grouping, determining which concepts to “keep.” For each phrase, we went sequentially through the list of concepts returned. The first concept in the list was automatically kept. If it had any words in common with the phrase, these words were flagged as having been “matched.” If there were any concepts with weight >= 11.0 (generally indicating a perfect match or near-perfect match), these concepts were also kept. Matched words in the phrase were again flagged. We then continued going down the list of concepts, looking to see if any concept contained a word or words that were still unmatched in the phrase. If so, we “kept” this concept as well, and flagged the matching word(s) as “matched.” Eventually we would run out of words to match in the phrase, or exhaust the list of concepts.

Our matching algorithm allowed words to differ by the letter “s” at the end — so “egg” matched to both “eggs” and “egg,” and “eggs” matched to both “eggs” and “egg.” The matching algorithm also ignored differences between uppercase and lowercase.

With the list of kept concepts, we removed any that were on the list of “stop concepts.” This was a list of 46 concepts, such as “Findings” and “History,” that were often returned but had little value in indexing. Finally, we ran the concepts through a “semantic filter.” That is, we checked the semantic type or types of each concept, and kept only the concepts that had a semantic type on the “keep list.” Some examples of kept semantic types are “Body Location or Region” and “Anatomical Abnormality.” Examples of undesired semantic types are “Spatial Concept” and “Qualitative Concept.” For those concepts that had more than one semantic type, if at least one of the types was on the “keep list,” the concept was kept.

The remaining concepts were the concepts selected by Chart Engine for the dictation. The entire process is diagrammed in Figure 2.

With the current version of SAPHIRE, our algorithm gave an overall precision of 0.241; for the 16-scan training set, the precision was 0.275 and for the 34-scan test set it was 0.228. The overall recall was 0.600; for the 16-scan training set it was 0.652 and for the 34-scan test set it was 0.579.

Research Question

We started by asking the question, “Could Natural Language Processing techniques improve the performance of our system?” The systems described above had all benefited to some degree from the use of NLP techniques.

We were using a list of barrier words to divide sentences into candidate phrases. The use of barrier words to divide sentences into phrases is not without precedent. This approach has shown some promise in the biomedical literature (Termsette et al., 1988), and as described in the recent paper by Nadkarni et al. (2001), continues to be used.

Our list of barrier words was derived from an original list containing 248 words. The words on the original list were, for the most part, prepositions (e.g., “in,” “of,” “from”), pronouns (e.g., “he,” “his,” “us”), articles (“an” and “the”) and other determiners (e.g., “this,” “these,” and “those,” “which,” “who,” and “what,” “many,” “any,” and “some”), conjunctions (e.g., “or,” “and”) and certain basic verbs (e.g., “is,” “are,” “be,” “has,” “have,” “did,” “do,” “can,” “could,” “might”). During the development of Chart Engine, other words, most notably verbs such as “show,” “shows,” “demonstrate,” and “demonstrated” were added to make a list of 290 barrier words. The final list with the added words flagged is given in Appendix B.

Candidate phrases were those words that fell between two barrier words, between a barrier word and certain punctuation (commas, periods, question marks, exclamation marks, colons, and semicolons), or between two punctuation marks. For example, the sentence

“Multiple areas of increased tracer uptake are seen in the lumbar spine, thoracic spine, ribs and bilateral sacral iliac joints.”

contained the barrier words “of,” “are,” “seen,” “in,” “the,” and “and.” Thus the phrases formed were “multiple areas,” “increased tracer uptake,” “lumbar spine,” “thoracic spine,” “ribs,” and “bilateral sacral iliac joints.”

During development of Chart Engine, several problems were noted with this barrier word approach. One problem was that using barrier words prevented proper interpretation of phrases containing conjunctions such as “and” or “or.” Take, for example, the string “flow void in the vertebral, basilar, and internal carotid arteries.” Using barrier words, this string was split into four candidate phrases: “flow void,” “vertebral,” “basilar,” and “internal carotid arteries.” However, the content of the string would have been more accurately represented by the candidate phrases “flow void,” “vertebral arteries,” “basilar arteries,” and “internal carotid arteries.” This particular problem naturally led to some failures to find indexing concepts (“recall failures”) when testing Chart Engine.

Another problem that we suspected led to recall failures was the fact that using barrier words sometimes excluded important trailing prepositional phrases. For example, the string “base of the skull” was divided into the phrases “base” and “skull” because “of” and “the” were barrier words. Yet neither “base” nor “skull” completely represented the concept expressed in the phrase “base of the skull.” Moreover, Chart Engine did not find the Metathesaurus concept “skull base” from either the phrase “skull” or the phrase “base” alone. In developing Chart Engine, we partly corrected for this problem for strings like “CT scan of the abdomen.” As described in the Chart Engine algorithm, this string was converted to “CT scan abdomen” within the sentence before the sentence was split up into phrases — thus the words “CT,” “scan,” and “abdomen” were sent as a single entity to SAPHIRE, giving a better chance that a concept like “Computerized axial tomography of the abdomen” would be found. Even so, we hypothesized that including *all* trailing prepositional phrases would further improve recall.

Furthermore, deficiencies in the list of barrier words sometimes caused two separate concepts to be lumped into one candidate phrase. For example, the string “adjacent cortical sulci causing effacement” did not contain any barrier words, and so was sent to SAPHIRE as a single phrase. In some cases this tended to give us spurious concepts — concepts that were not desired for indexing, also known as “precision failures.” Again, we partly corrected for this problem in the development of Chart Engine by adding derivatives of common verbs like “to show” and “to demonstrate” to our barrier word list. However, we also hypothesized that correcting this problem more definitively would eliminate additional errors in precision.

Finally, Chart Engine missed some indexing concepts because the candidate phrases it created contained extra descriptors. The Metathesaurus concept “uptake,” for example, was present as an indexing concept in some of the documents. However, if the word “uptake” occurred in a string like “increased tracer uptake,” the concept “uptake” would not be found. However, the concept could be found from the truncated phrase “tracer uptake.” In a similar fashion, the Metathesaurus concept “abdomen” was not found from the candidate phrase “upper abdomen,” but could be found from the shorter phrase “abdomen.”

These problems with phrases derived from barrier words led us to try other methods of choosing candidate phrases. We used an evaluation copy of NPtool, a commercial noun-phrase extractor (product of Lingsoft, www.lingsoft.fi), to select candidate phrases to send to SAPHIRE. This led to an improvement in recall (0.735), but a decrease in precision (0.160). However, these candidate phrases were derived from documents from which historical sentences were not removed and negated and normal phrases were still sent as candidate phrases. In comparison, with our barrier word approach we obtained a recall of 0.617 and a precision of 0.196 on documents in which no negated phrases or historical sentences were removed. But because NPtool outputted only noun phrases without any context, it was difficult to tell which noun phrases were negated (e.g., "chest nodule" in the sentence "There is no chest nodule.") or normal (e.g., "lungs" in the sentence "The lungs are clear.").

A competitor of Lingsoft offered a tool that provided part-of-speech tagging for sentences. We licensed this tool, called EngLite (product of Conexor oy, www.conexor.fi), and hypothesized that we could use this tool to create candidate phrases that would overcome the limitations of phrases formed by the barrier word technique. In particular, we hypothesized that the NLP capabilities that this tool provided would give us the flexibility to define candidate phrases *and* to incorporate recognition of negated and normal phrases in such a way to improve precision and recall.

Materials and Methods

Syntactic Parsing with EngLite

For this project we used EngLite to send candidate phrases to SAPHIRE. EngLite took English text and assigned each word:

1. a root form, or “lemma,” and
2. an interpretation of the word as a component of the sentence, usually consisting of a syntactic tag and part of speech.

Sometimes more than one interpretation of the word was given, resulting in additional fields.

As a parser, EngLite processed text automatically. Being a syntactic parser, it used information about English syntax to process the text. In particular, it used a *lexicon* containing information about English words and their structure as well as a set of syntactic rules. It also featured a guesser that attempted to assign correct parts of speech to words that were not in the lexicon.

EngLite is a shallow syntactic parser. While it functioned to mark off noun phrases, it did not give information such as whether a noun phrase functioned as a *subject* (the actor) or *object* (the one acted upon) in a sentence. For example, the phrases “the man” and “the dog” would receive the same interpretation in the sentence “The man bit the dog.” Again, the phrases “the star,” “the jar,” and “the bar” would all receive the same interpretation in the sentence “I saw the star in the jar at the bar.”

A few further examples help to show how EngLite functions. The following four phrases are all noun phrases:

“Bill”

“the water boy”

“the girl with blue eyes”

“Oregon Health & Science University”

.

EngLite output, below, from the word “Bill” affirms that its root form is “Bill,” its syntactic function is as a noun head, and its part of speech is noun. The <s> on the next line denotes that EngLite recognized the end of the phrase as the end of a sentence.

```
Bill Bill &NH N
<s>  <s>
```

EngLite output from the phrase “the water boy” shows that “the,” “water,” and “boy” are already in their root forms. Furthermore, “the” is a determiner (DET) and functions to modify a noun head that comes after it (&>N). “Water” is a noun (N) and also functions to modify a noun head that comes after it (&>N). “Boy” is a noun (N) and is the noun head (&NH) of this phrase.

```
the  the  &>N DET
water     water     &>N N
boy  boy  &NH N
```

```
<s>  <s>
```

In the output from the phrase “the girl with blue eyes,” the root forms are all the same as the original word, except for “eyes,” whose root is “eye.” The word “the” functions as it did in the last example, as a determiner (DET) that modifies (&>N) the noun, “girl,” that comes after it. “Girl” functions as a noun (N) and noun head (&NH). “With” is a preposition (PREP) that comes after “girl,” and as part of the prepositional phrase “with blue eyes,” modifies it (&N<). “Blue” is an adjective (A) occurring locally to modify the noun head “eyes.” “Eyes” is a noun and noun head. Note that even though the whole phrase “with blue eyes” modifies “the girl,” only the word “with” has a syntactic tag that indicates that it modifies a noun before it. Indeed, the string “blue eyes” forms a separate noun phrase.

```
the     the     &>N DET
girl    girl    &NH N
with    with    &N< PREP
blue    blue    &>N A
eyes    eye     &NH N
<s>     <s>
```

In the last example, output from the phrase “Oregon Health & Science University,” the root forms are all the same as the original words, ignoring capitalization. The symbol “&,” standing for the word “and,” is denoted simply as a “&CC,” or coordinating conjunction. “Oregon” is a noun (N) that modifies the noun head “University” (&>N).

However, in this phrase, “health” can take one of two meanings. It is either a noun (N) that is a noun head (&NH), or a noun (N) that modifies a succeeding noun head (&>N). “Science” is a noun (N) that modifies a succeeding noun head (&>N), and “university” is a noun (N) that functions as a noun head (&NH).

```
Oregon  Oregon  &>N N
Health  health  &>N N   &NH N
&       &       &CC
Science science &>N N
University      university      &NH N
<s>     <s>
```

The two meanings of “health” reflect uncertainty on the part of the parser. This in turn reflects syntactic ambiguity inherent in the sentence.

A complete listing of the part-of-speech and syntactic tags used by EngLite is given in Appendix C. Allen (1995) offers a review on English grammar that covers all of the concepts above.

Mapping to the Metathesaurus with SAPHIRE

SAPHIRE is a computer application originally developed by Hersh and Greenes (Hersh and Greenes, 1990). Its function is to map chunks of text to Metathesaurus concepts. It does so by exploiting the fact that the Metathesaurus contains a great deal of information about synonyms because it contains synonyms that are linked from many diverse medical vocabularies. For example, the same concept is known as “Cervix uteri” in the 2000 version of the Medical Subject Headings and “Uterine cervix” in SNOMED version 2. Its canonical, or preferred form, is “Uterine cervix-Anatomy,” with Concept Unique Identifier or CUI “C0007874.”

The current version of SAPHIRE matches Metathesaurus concepts to text by a ranked, partial-matching process. For a given input of text, a list of concepts is returned, each concept having a *weight*. In general, the higher the weight, the better a match the concept is to the text. SAPHIRE functions with a default minimum cutoff weight of 1.0. The usual maximum weight, 11.0 or greater, represents a perfect or near-perfect match between input text and concept. SAPHIRE, for example, gives the following output for the input “heart attack”:

```
C0027051 1.693147 Myocardial Infarction
C0155668 1.399075 Old myocardial infarction
C0497237 1.193147 Fear of heart attack
C0018787 1.000000 Heart
C0153500 1.000000 malignant neoplasm of heart
C0153957 1.000000 benign neoplasm of heart
C0699795 1.000000 Attack
C0795691 1.000000 HEART PROBLEM
```

Here, the first column is the concept unique identifier or CUI, the second column is the SAPHIRE weight, and the third column is the canonical or preferred name of the concept. In this example "Myocardial Infarction" (CUI C0027051), the top concept returned, is indeed the best match to the text "heart attack."

Programming and Data Analysis

Programming for this project was done in PERL 5 on a Sun Solaris server running Solaris 2.7. A suite of programs, Phrase Generator, were built to take the output from EngLite and construct phrases from it. The programs were built based on an algorithm given in Allen (1995), outlined in Appendix D and explained below.

Taking the EngLite output of one English sentence at a time, we used Phrase Generator to form phrases according to a set of rules that could be easily changed. We encoded each rule as a symbolic expression. One rule, for example, stated that a noun phrase could be a noun head. This rule could be written as

1. n_p ← n_h

Elsewhere in the suite of programs, we explicitly defined a noun head, or "n_h" as *any* word to which EngLite gave the syntactic tag "&NH." Thus the words "Bill," "boy," "girl," "eyes," "Health," and "University," from the four sample noun phrases given above, were "n_h"s.

In a similar fashion, we wrote other rules allowing noun phrases to consist of a sequence of "pre-nouns" followed by a single noun head. A "pre-noun" (p_n) was defined as a noun (N) that had the syntactic tag “&>N,” indicating that it modified a noun head following it. Thus the words “water,” “Oregon,” “Health,” and “Science” would all be pre-nouns. Allowing pre-nouns to come before a noun head in a sentence could be specified by the following three rules:

2. n_p ← n_p2
3. n_p2 ← n_h
4. n_p2 ← p_n n_p2.

To process these types of rules, which are known as a *Context Free Grammar*, our Phrase Generator used an algorithm called “bottom-up chart parsing.” (Allen, 1995)

The idea of a bottom-up Chart Parser is to start at the beginning of a string of words or sentence and classify each word in turn. As each word is classified, information is added to the “chart”: “word #1 is a p_n, word #2 is a n_h.” The chart parser attempts to “build” bigger pieces from smaller ones, until it cannot build any more.

The best way to illustrate how a bottom-up chart parser words is with an example. Take the string “house boat,” and assume that EngLite has tagged “house” as a p_n (pre-noun) and “boat” as an n_h (noun head).

The chart parser first looks in the agenda, which is a sort of “to do” list of items that have been classified but not fully added to the chart. Since we have just started the sentence, the agenda is empty. So the parser looks at the first word, “house.” It confirms that "house" is a p_n, and adds information to the effect of "Word #1 in the string is a p_n" to the chart. Then it looks to see if "house," as a p_n, triggers any rules. The right-hand side of Rule #4, n_p2 ← p_n n_h, begins with a p_n. So the parser adds an "active arc" to the chart, stating in effect that "Word #1 matches the first part of Rule #4, and we're still waiting for an n_p2 to complete this rule."

Next the chart parser looks at the second word, "boat" (the agenda is still empty). It confirms that "boat" is an n_h, and adds the information "Word #2 is an n_h" to the chart. It looks to see if "boat" as an n_h triggers any rules. Two rules are triggered: rule #3 and rule #1. Rule #3, "n_p2 ← n_h," is *completed* by "boat" (if we've found an n_h, we must have found an n_p2). So a new component is added to the chart, stating that "Words #2 forms an n_2." In addition, this component is added to the agenda, to be processed in turn. Similarly, Rule #1, "n_p ← n_h," is completed by "boat," and so the component stating that "Word #2 forms an n_p" is added to the agenda. Finally, looking at the set of active arcs, the parser determines that "boat" completes the active arc started by "house." Since the rule for this arc stated that "n_p2 ← p_n n_h," the parser adds the component stating that "Words #1 and #2 form an n_p2" to the agenda.

All of this is wonderfully tedious and slow, but in the end the parser recognizes that "house" is a p_n, "boat" is an n_h, since "boat" is an n_h it is also an n_p and an n_p2, and that "house boat" is an n_p2. Finally, since "house boat" is an n_p2 it is also an n_p.

Having built this parser, we were able to set up rules for what constitutes a noun phrase. We had great flexibility with these rules. Most of all, the parser allowed us to handle the situation where a word could be classified in two separate ways — say as a pre-noun and as a noun head. The parser could then form noun phrases using either interpretation, often equivalent to giving us two conflicting interpretations of the sentence.

We created two specific sets of rules, corresponding to the notions of "simple noun phrases" and "noun phrases with prepositional attachments," respectively (Appendix E). The first rule stated that "a noun phrase is a sequence of zero, one, or two determiners, followed by an arbitrary number of adjectives, followed by an arbitrary number of *pre-nouns*, all modifying a single noun head." A determiner was defined as any word that received the part-of-speech tag DET. Examples of determiners include "the" in "the big man," or "these" in "these days." An adjective was defined as a word that received the syntactic tag &>N and *did not* receive the part-of-speech tag N. As in the explanation above, a pre-noun was defined as a word that received the syntactic tag &>N and the part-of-speech tag N (for noun), and a noun head was defined as any word that received the syntactic tag &NH, regardless of part-of-speech tag.

The second rule stated that "a noun phrase is an arbitrarily long sequence of noun phrases allowed by the first rule, provided each noun phrase is separated from the next by exactly one preposition that has the syntactic tag &N<." In other words, one could add successive prepositional phrases to the end of noun phrases, as long as each preposition contained the tag “&N<” to indicate that the prepositional phrase modified the noun coming before it.

The first set of rules followed fairly logically from the syntactic tags given to us by EngLite: “&>N,” a pre-modifier of a noun head, and “&NH,” a noun head. In our rules, however, an assumption was made that pre-modifiers would always occur in the sequence “determiner, adjectives, pre-nouns.” This was in fact not always the case. The second set of rules also follows in a straightforward fashion from EngLite’s syntactic tags.

After the noun phrase generator was successfully built and debugged, the ability to recognize negated noun phrases was programmed. Using the negation algorithm from the baseline Chart Engine algorithm, a noun phrase was assumed to be negated if it was a substring of some string that would have been negated under the old system. Negation was tested and debugged.

Finally, the ability to recognize “maximal simple noun phrases” was programmed. The PERL scripts were enabled to determine if a phrase from a sentence was “maximal,” that is, that there were no other noun phrases in that sentence that shared its terminal word and had extra determiners, adjectives or pre-modifiers in the beginning.

Using these tools and rules, three sets of noun phrases, Run 1, Run 2, and Run 3 were generated from the entire set of 50 dictations and run through the standard Chart Engine algorithm. Runs 1 and 2 were derived from the first set of rules. Run 1 contained simple noun phrases that were maximal. Run 2 contained *all* simple noun phrases. Run 3 was derived from the second set of rules, and contained all simple noun phrases and noun phrases with prepositional attachments.

The phrases and results using the barrier word list were designated as Run 0. By definition, the candidate phrases in Run 0 were derived in a completely different manner from the candidate phrases in Runs 1, 2 and 3. The phrases in Run 1, however, were a subset of the phrases in Run 2, and the phrases in Run 2 were a subset of the phrases in Run 3. Considering as an example the sentence:
“The sun in the blue sky is beautiful.”

The barrier words in this sentence are “the,” “in,” and “is.” So Run 0, the baseline run, would have contained the phrases “sun,” “blue sky,” and “beautiful.”

Runs 1, 2 and 3, on the other hand, would have been constructed from the EngLite parse of the sentence as shown below:

```
The     the     &>N DET
sun     sun     &NH N
in      in      &N< PREP
the     the     &>N DET
blue    blue    &>N A
sky     sky     &NH N
```

```
is      be      &VA V
beautiful       beautiful       &NH A
.       .

<s>     <s>
```

According to the rules for producing noun phrases, then, Run 1 would contain the phrases “the sun” (a determiner plus a noun head), “the blue sky” (a determiner plus adjective plus noun head), and “beautiful” (a noun head). Run 2 would contain the same phrases as Run 1, plus the phrase “sun” (derived from the phrase “the sun”) and the phrases “blue sky” and “sky” (derived from the phrase “the blue sky”). Run 3 would contain all of the phrases contained in Run 2, plus the additional phrases “the sun in the blue sky” and “sun in the blue sky” (two noun phrases separated by a preposition that is a noun post-modifier).

Precision and recall were computed automatically for each of the runs as defined above. We compared precision and recall between the baseline and experimental runs, and then investigated the differences in a qualitative fashion.

Results

Table 1 shows computed precision and recall for the four runs. In general, recall increased with successive runs while precision decreased.

Table 2 describes the number and length of unique candidate phrases generated during each run. Figures for Table 2 were obtained by pooling all of the candidate phrases generated from the 50 scans into a single file, converting uppercase letters to lowercase, and removing duplicates. A "word" is defined as a sequence of characters separated by whitespace, according to the UNIX wc utility. Run 0 has only 1354 unique phrases, while Run 1 has 1522, Run 2 has 2712, and Run 3 has 4414. These figures are consistent with the definition of the runs.

	Concepts Found Matching Gold Standard	Total Concepts In Gold Standard	All Concepts Found	Precision	Recall
Run 0	299	499	1242	0.240741	0.599198
Run 1	294	499	1223	0.240392	0.589178
Run 2	348	499	1604	0.216958	0.697395
Run 3	351	499	1852	0.189525	0.703407

Table 1: Precision and Recall for Runs 0, 1, 2, and 3

Run	Number of unique phrases	Average number of words per phrase	Longest phrase (words)
Run 0	1354	2.428	11
Run 1	1522	2.489	8
Run 2	2712	2.333	8
Run 3	4414	3.844	19

Table 2: Characteristics of phrases for Runs 0, 1, 2, 3

Run 0 vs. Run 1

Despite the difference in the way the phrases were generated, Run 0 and Run 1 produced markedly similar results, finding 291 concepts in common. Only 11 concepts (3.6%) were found in one run but not another.

A total of eight concepts were found in Run 0 but not in Run 1. Two of these differences had to do with the way the text was parsed. First, in Chest CT #7, the concept “Computerized tomography guided biopsy” was found in Run 0 from the phrase “CT guided biopsy.” However, in Run 1 this concept was not found because EngLite parsed “CT guided biopsy” erroneously as:

```
CT  CT  &NH N
GUIDED  guide   &VA V
BIOPSY  biopsy  &NH N
.   .

<s> <s>
```

Likewise, in Head CT #1, the concept “CT of head” was found in Run 0 from the phrase “head ct.” However, in Run 1 EngLite erroneously parsed “head ct” as

```
HEAD    head    &VA V
CT  CT  &NH N
.   .
```

```
<s> <s>
```

It did not recognize that “head” modified “ct”; thus the concept “CT of head” was not found.

A further two concepts were not found in Run 1 because the “gold standard” word was not a noun phrase. In Head CT #7, the words “calcified” and “compressed” were both interpreted by EngLite as a past participle of a verb (EN) functioning as a passive verb (&VP). While one might argue with these interpretations, the strings “calcified” and “compressed” are clearly not noun phrases, in that neither word can be a noun.

```
This    this    &>N DET
mural   mural   &>N N
nodule  nodule  &NH N
is  be  &AUX V
partially   partial &AH ADV
calcified   calcify &VP EN
```

```
The the &>N DET
ipsilateral ipsilateral &>N A
temporal    temporal    &>N A
horn    horn    &NH N
is  be  &AUX V
also    also    &AH ADV
```

```
compressed  compress    &VP EN
```

The other four concepts were not found in Run 1 because EngLite did not recognize forward slashes as boundaries between words, while the barrier word method did. These concepts were “pelvis” from the text “ABDOMEN/PELVIS” in Abdominal CT #9, “edematous” and “reaction” from the string “mass effect/edematous/inflammatory reaction” in Head MRI #3, and “left axilla” from the text “left axilla/left anterior chest wall” in Abdominal CT #7.

Three concepts were found in Run 1 but not in Run 0. The concept “Uptake” was not found in Run 0 in Bone Scan #2 because it appeared everywhere as “increased uptake.” However, in Run 1, “increased uptake” was parsed differently in two places in the radiology dictation. In the dictation body, “increased uptake” appears within the context of “Increased uptake in the left 7th rib approximately is noted” and was parsed as

```
Increased increase  &>N EN
uptake    uptake    &NH N
in   in   &N< PREP
the  the  &>N DET
anterior  anterior  &>N A
aspect    aspect    &NH N
of   of   &N< PREP
the  the  &>N DET
left left &>N A
7th  7th  &>N NUM
```

```
rib  rib  &NH N
approximately  approximate    &AH ADV
is   be   &AUX V
noted     note &VP EN
.    .

<s>  <s>
```

In the Impression section of the dictation, however, the telegraphic sentence "INCREASED UPTAKE IN THE LEFT RIB AS DESCRIBED ABOVE, LIKELY TRAUMATIC" was parsed as follows:

```
INCREASED   increase    &VA V
UPTAKE  uptake  &NH N
IN  in  &N< PREP
THE the &>N DET
LEFT    left    &>N A
RIB rib &NH N
AS  as  &CS
DESCRIBED   describe    &VP EN
ABOVE   above   &AH ADV
,   ,
LIKELY  likely  &>A ADV &AH ADV &>N A
TRAUMATIC   traumatic   &NH A
.   .
```

```
<s> <s>
```

The second time, only the word “uptake” was recognized as a noun phrase — the word “increased” was tagged as being a verb (V) that functions in the sentence as an active verb (&VA).

Likewise, the concept “basal ganglia” in Head CT #7 was not found in Run 0 because it appeared as the phrase “right basal ganglia” or as “large right basal ganglia mass extending.” However, in Run 1, “large right basal ganglia mass extending…” was parsed as

```
LARGE   large   &>N A   &NH A
RIGHT   right   &>A ADV &AH ADV &>N N
BASAL   basal   &>N A
GANGLIA ganglion    &NH N
MASS    Mass    &>N N
EXTENDING   extend  &NH ING
```

Thus “basal ganglia” appeared as a maximal simple noun phrase because “right” was interpreted as a possible pre-noun, and “basal” was interpreted as an adjective.

The final concept found in Run 1 but not in Run 0 was due to a fluke — two errors. The relevant sentence in Abdominal CT #8 read:

“No enhancing lesions in the cerebral hemispheres, basal ganglia, or cerebellar hemispheres.”

However, due to an apparent error on the indexer’s part, the Metathesaurus concept “basal ganglia” was chosen as an indexing term. In Run 0, the phrases “enhancing lesions,” “cerebral hemispheres,” “basal ganglia,” and “cerebellar hemispheres” were not sent as candidate phrases because of our negation algorithm. But in Run 1, the two candidate phrases “basal ganglia” and “cerebellar hemispheres” were still sent due to a bug in the negation algorithm in adapting it from the barrier word method of generating phrases to the new method. (This type of error, where only the first phrase in a comma-separated list of negated phrases is eliminated, occurred in several other places as well). However in this case, the candidate phrase that was negated turned out to yield a gold standard concept, “Basal ganglia.”

Run 1 vs. Run 2

As noted above, the phrases in Run 2 (all noun phrases derived from maximal simple noun phrases that contained the noun head) were a superset of the phrases in Run 1 (all maximal simple noun phrases). There were an additional 1190 unique phrases generated from the 50 scans in Run 2 as compared to Run 1.

In Run 2 an additional 54 gold standard concepts were found. For each of these 54 additional concepts, the successful candidate phrases in Run 2 were, by definition, subphrases of unsuccessful phrases in Run 1. For example, the phrase “increased tracer uptake” in Run 1 (Bone Scan #1) was unsuccessful in finding the gold standard concept “uptake.” These following lines show the phrase, “increased tracer uptake,” and the three concepts that our algorithm finds for that phrase. Each concept is listed with its CUI, SAPHIRE weight, canonical form, source vocabulary or vocabularies, and semantic type(s).

```
increased tracer uptake|
    C0205217|1.000000|Increased|RCD99   SNMI98  |Functional
Concept
    C0241323|1.000000|T3 UPTAKE, INCREASED|DXP94   |Finding
    C0684208|1.000000|tracer|PDQ99   |Indicator, Reagent,
or Diagnostic Aid
```

However, the subphrase “tracer uptake” in Run 2 was successful:

```
tracer uptake|
    C0243144|1.000000|uptake|CSP98   |Physiologic Function
    C0684208|1.000000|tracer|PDQ99   |Indicator, Reagent,
or Diagnostic Aid
```

Likewise, the phrase “whole body bone scan” in Run 1 (Bone Scan #1) was unsuccessful in finding the concept Radioisotope scan of bone, NOS:

```
WHOLE BODY BONE SCAN|
    C0203669|2.098612|Total body scan|ICD2000 SNMI98
|Diagnostic Procedure
    C0037884|1.693147|Sphenoid Bone|MSH2000 |Body Part,
Organ, or Organ Component
```

However, the subphrase “bone scan” in Run 2 was successful:

```
BONE SCAN|
    C0203668|1.693147|Radioisotope scan of bone, NOS|SNMI98
|Diagnostic Procedure
```

Again, in Chest CT #2, Run 1 contained the candidate phrase “the left kidney,” which gave the following Chart Engine output:

```
the left kidney|
    the left kidney|C0227614|11.693147|Left kidney|RCD99
SNM2    SNMI98  UWDA99  |Body Part, Organ, or Organ
Component
```

Run 2 contained the subphrase “kidney” as well, which gave the desired indexing concept, “Kidney”:

```
kidney|
    kidney|C0022646|11.000000|Kidney|ICD10   MSH2000 MTH
PCDS97  RCD99   SNM2    ULT93   UWDA99  |Body Part, Organ,
or Organ Component
```

In general, a median of two words had to be removed from unsuccessful phrases in Run 1 before getting successful phrases in Run 2. At one extreme, the phrase “left anterior descending coronary artery” in Run 1 was assigned the indexing concept “Coronary artery, NOS.” Run 2 contained the subphrases “anterior descending coronary artery,” “descending coronary artery,” “coronary artery” and “artery” as well as the full phrase. Only the phrase “coronary artery,” when processed by SAPHIRE and our post-processing algorithm yielded the indexing concept “Coronary artery, NOS.”

Further analysis of the differences between Run 1 and Run 2 shows that having the words “the” or “a” at the beginning of a phrase did not lead to any recall successes. In fact, if the modifiers “left,” “right,” “the left,” and “the right” had been removed from phrases in Run 1, an additional 18 concepts would have been found. Not taking into account the articles “the” and “a,” an additional 38 concepts in all would have been found if a single modifier had been removed from the front of a phrase in Run 1. Most often these modifiers consisted of the following:

- The words “right” or “left.”
- Spatial concepts (“upper,” “distal”).
- Anatomic concepts (“pulmonary,” “lumbar”).
- Qualitative concepts (“benign,” “extensive”).

49 extra indexing concepts would have been found by making additional phrases by systematically removing one word and two words from all of the phrases in Run 1.

Run 2 vs. Run 3

The phrases in Run 3 (all derivatives of noun phrases including prepositional attachments) are a superset of the phrases in Run 2 (all derivatives of simple noun phrases). There are an additional 1702 unique phrases in Run 3 as compared to Run 2. However, only three extra concepts are found in Run 3 that are not found in Run 2:

- The concept “right lobe of liver, NOS” from “right lobe of the liver” in Abdominal CT #4,
- The concept “computerized axial tomography of thorax without contrast” from “contrast with axial 3 mm” in Abdominal CT #8, and
- The concept “skull base” from “base of skull” in Head CT #9.

At the same time, precision decreased slightly from Run 2 to Run 3 (from 0.217 to 0.190). A total of 245 unnecessary concepts were found in addition to the three gold standard concepts found.

Recall Failure Analysis

Over the 50-scan collection, there were 205 indexing concepts not found in Run 1, and 151 indexing concepts not found in Run 2. Again, if a concept, e.g., “base of skull,” was designated as an indexing concept for a certain scan, say Head CT 1, then we only counted the indexing concept as “found” if this concept was found from phrases derived from that specific scan. Even if “base of skull” was a concept found from Head CT 2, if it was not found from Head CT 1, we counted that indexing concept to have been missed.

As noted before, the set of indexing concepts found did not differ substantially from Run 0 to Run 1.

The set of 151 indexing concepts not found in Run 2 was analyzed and classified. We denoted each of these omissions “recall errors” because each indexing concept not found had a direct impact on calculated recall, more so than precision. Of these 151 recall errors:

- 23 recall errors were indexing concepts that were filtered by semantic type. They were returned by SAPHIRE, and retained in initial post-processing, but discarded because their semantic type was not listed as desirable.
- 12 recall errors were indexing concepts that could not be found because of coordination problems. For example, the phrase “vertebral, basilar, and internal carotid arteries” was assigned the indexing concepts “vertebral artery,” “basilar artery” and “internal carotid artery.” Our noun phrase-making algorithms, however, were not programmed to be able to generate the phrases “vertebral arteries” and “basilar arteries.”
- 11 recall errors were indexing concepts that corresponded to a complex noun phrase, i.e. one with a prepositional phrase. For example, the indexing concept “base of

skull" was represented in the text with the words "base of skull," i.e. a noun phrase containing a prepositional phrase.

- 4 concepts were not found because our algorithm designated the sentences in which they occurred "historical sentences" and eliminated them from further consideration.
- Similarly, 7 concepts were not found because potential candidate phrases were designated as negated (5) or normal (2) and eliminated.
- 59 other concepts were potentially findable, but not found.
  - For 14 of these concepts, we did not send off any appropriate candidate phrase
  - For the other 45 concepts, we sent off a candidate phrase similar in meaning to the indexing concept, but did not end up either returning or keeping the indexing concept
- 35 concepts were apparently not findable in the current version of the Metathesaurus. That is, the dictations were indexed with concepts from the 1998 or 1999 version of the Metathesaurus. However, we used an algorithm that returned concepts from the 2000 version of the Metathesaurus. In 35 instances, an indexing concept had been assigned a different CUI in the 2000 Metathesaurus or had been removed from the Metathesaurus.

Precision Analysis

There were 943 precision errors in Run 0, 929 in Run 1, 1256 in Run 2, and 1501 in Run 3. The details of these errors were not analyzed.

Discussion

Unexpectedly, there was not much difference between the results taking maximal simple noun phrases generated using a shallow syntactic tagger (Run 1) and the baseline results, using phrases generated with a barrier list (Run 0). There are a few reasons why this might be the case. Run-on phrases (“adjacent cortical sulci causing effacement”) occurred infrequently — less than once per dictation — and so we probably realized less benefit than expected by breaking up these phrases. The inclusion of the definite article “the” in Run 1 phrases caused some precision errors, while conferring no benefit on recall. Probably most importantly, the Chart Engine algorithm had been tuned for peculiarities of the barrier word phrase-maker (e.g., the fact that it separated tokens at a forward slash), while it was not tuned for the phrases generated in this project.

The problem of conjunctions, posed in the introduction, proved to be out of the scope of this work. Correctly subdividing phrases with conjunctions is not a trivial problem, and a definitive solution must make use of semantic and pragmatic knowledge. For example, the string “vertebral and internal carotid arteries” is syntactically ambiguous between several interpretations, including “vertebral artery and internal carotid artery” and “vertebral carotid artery and internal carotid artery” (clearly false). In any event, we would be unlikely to realize a substantial benefit from incorporating coordination, as only 11 concepts out of 499 were not found due to its absence.

As hypothesized, a big benefit in recall, with a decrease in precision, was found when all subphrases from the maximal noun phrases (Run 2) were taken. The apparent reason was that pre-modifying adjectives and/or other parts of speech interfered with SAPHIRE’s recognition of key concepts. A systematic analysis could be done of potential modifiers that could be eliminated, so that recall could be increased but precision maintained or even increased. For example, the adjective “numerous” did not seem to add any benefit when added to phrases. However, like using lists of barrier words, using lists of modifiers to remove has the potential to degrade recall — consider if we were to remove “multiple” from the phrase “multiple myeloma,” or “left” from “(persistent) left superior vena cava.”

We saw only a slight increase in recall when complex noun phrases were added, and a further decrease in precision. This result was somewhat surprising, and may in part reflect grammatical usage in our dictations or in radiology dictations in general. There may be a tendency for radiologists in dictation to use the shortened forms of phrases (“bone marrow cancer”) as opposed to the longer forms (“cancer of the bone marrow”). One might expect this in dictation, where speed is often of the essence. This phenomenon may also reflect the fact that the concepts we were required to find were more often simple (“skull”) rather than complex (“tumor of the skull”). Furthermore, we had already corrected for the fact that procedural concepts (“CT scan of the head”) consisted of complex phrases in our baseline algorithm by getting rid of the preposition and article in our preprocessor for phrases of the form “CT [scan] of [the]” … or “MRI [scan] of [the]….” Nevertheless, the failure analysis of Run 1 reveals 11 gold standard concepts whose meaning was contained in a phrase split across a preposition; yet we found only three of these concepts by including trailing prepositional phrases to our candidate phrases that we sent through Chart Engine. It is unclear what to make of this finding.

Limitations

There were several factors that limited the internal validity of this study. First, only a single medical librarian did the indexing. A study of trained professionals indexing MEDLINE documents has demonstrated that the reproducibility of MEDLINE indexing is limited (Funk and Reid, 1983). Reproducibility would likely also be limited among professionals indexing radiology dictations. Second, the set of dictations we used was small and came from only one institution. Dictations at another institution or even from different radiologists at the same institution may reflect different language and may have produced different results. Third, the Chart Engine algorithm built on SAPHIRE was finely tuned to the data set and to the phrases formed using barrier words. It may be that results would improve following tuning for phrases formed using the syntactic methods.

In a broader sense, the study is limited as to what it can say about the effectiveness of using syntactic techniques for forming noun phrases for indexing as compared to the effectiveness of using barrier words. We approached the task of indexing medical text by assigning concepts to the text from a standardized vocabulary. There are several other approaches, including indexing the text by word-statistical methods and indexing it using statistically generated phrases. Again, in the medical domain there are many standardized vocabularies available. We studied only one, the UMLS Metathesaurus. Finally, there are a number of tools to map from text to standardized vocabularies. We chose to look at one, SAPHIRE, and the Chart Engine algorithm built using SAPHIRE.

Nevertheless, this study provides more evidence that the use of barrier words as a "poor man’s parser” may be good enough for some information retrieval tasks. Were we to want near-perfect precision and recall, neither our use of barrier words nor our use of limited syntactic techniques would suffice. But for some of the tasks mentioned in the introduction — assembling a teaching file of chest x-rays for residents, for instance — barrier word techniques may be “good enough.”

Summary and Conclusions

In summary, we built a flexible noun phrase recognizer atop a shallow syntactic parser and substituted its output for phrases formed using a barrier word list in a concept-matching algorithm. The noun phrase recognizer appeared to function as effectively as the barrier word list, and we were able to incorporate negation with only a few minor errors. However, more time and effort would be necessary to turn this technique into something that could improve both the precision *and* recall of our system.

References

Allen J. Natural Language Understanding – Second Edition. The Benjamin/Cummings Publishing Company: Redwood City, CA, 1995.

Aronson AR. The effect of textual variation on concept based information retrieval. Annual Symposium of the American Medical Informatics Association, 1996, pp. 373-377.

Friedman C. Towards a comprehensive medical language processing system: methods and issues. Proc AMIA Annu Fall Symp. 1997:595-599.

Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Informatics Assoc. 1994;1:161-174.

Funk ME, Reid CA. Indexing consistency in MEDLINE. Bull Med Libr Assoc. 1983;71:176-183.

Gundersen ML, Haug PJ, Pryor TA, van Bree R, Koehler S, Bauer K, Clemons B. Development and evaluation of a computerized admission diagnoses encoding system. Compu Biomed Res. 1996;29:351-372.

Haug PJ, Ranum DL, Frederick PR. Computerized extraction of coded findings from free-text radiologic reports: work in progress. Radiology 1990;174:543-548.

Hersh WR. Information Retrieval: A Health Care Perspective. New York: Springer-Verlag, 1996.

Hersh WR, Greenes RA. SAPHIRE – an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. Compu Biomed Res. 1990;23:410-425.

Hersh W, Leone TJ. The SAPHIRE server: a new algorithm and implementation. Proc 19th Annu Symp Comput Appl Med Care. 1995:858-862.

Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med. 1995;122:681-688.

Hripcsak G. Knirsch CA, Jain NL, Pablos-Mendez A. Automated tuberculosis detection. J Am Med Informatics Assoc. 1997;4:376-381.

Lin R, Lenert L, Middleton B, Shiffman S. A free-text processing system to capture physical findings: Canonical Phrase Identification System (CAPIS). Proc 16th Annu Symp Comput Appl Med Care. 1992:843-847.

Lindberg DAB, Humphreys BL, McCray AT. The unified medical language system. Meth Inform Med. 1993; 32:281-91.

Lowe HJ, Antipov I, Hersh W, Smith CA. Towards knowledge-based retrieval of medical images. The role of semantic indexing, image content representation and knowledge-based retrieval. Proc AMIA Annu Fall Symp. 1998:882-886.

Lowe HJ, Walker WK, Vries JK. Using agent-based technology to create a cost-effective, integrated, multimedia view of the electronic medical record. Proc 19th Annu Symp Comput Appl Med Care. 1995:441-444.

Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases: a feasibility study. J Am Med Informatics Assoc. 2001;8:80-91.

National Library of Medicine. UMLS Knowledge Sources, 11th Edition. 2000.

Nelson SJ, Cole WG, Tuttle MS, Olson NE, Sheretz DD. Recognizing new medical knowledge computationally. Proc 18th Annu Symp Comput Appl Med Care. 1993;409-413.

Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Informatics Assoc. 1994;1:142-160.

Sager N, Lyman M, Tick LJ, Nhan TN, Bucknall CE. Natural language processing of asthma discharge summaries for the monitoring of patient care. Proc 18th Annu Symp Comput Appl Med Care. 1993:265-268.

Termsette KWF, Scott AF, Moore GW, Matheson NW, Miller RE. Barrier word method for detecting molecular biology multiple word terms. Proc 12th Annu Symp Comput Appl Med Care. 1988;207-211.

Tuttle MS, Olson NE, Keck KD et al. Metaphrase: an aid to clinical conceptualization and formalization of patient problems in healthcare enterprises. Meth Inform Med. 1998;37:373-83.

Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Computers and Biomedical Research 1993;26:467-481.

Background Reading

Baud R. Present and future trends with NLP. International Journal of Medical Informatics 1998;52:133-139.

Evans DA, Hersh WR, Monarch IA, Lefferts RG, Handerson SK. Automatic indexing of abstracts via natural-language processing using a simple thesaurus. Med Decis Making 1991;11(suppl):S108-S115.

Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. Meth Inform Med 1998;37:334-44.

Hersh WR, Campbell EM, Malveau SE. Assessing the feasibility of large-scale natural language processing in a corpus of ordinary medical records: a lexical analysis. Proc AMIA Annu Fall Symp. 1997:580-584.

Lowe HJ, Buchanan BG, Cooper GF, Vries JK. Building a medical multimedia database system to integrate clinical information: an application of high-performance computing and communications technology. Bull Med Libr Assoc 1995a;83(1):57-64.

Spackman KA, Hersh WR. Recognizing noun phrases in medical discharge summaries: an evaluation of two natural language parsers. Proc AMIA Annu Fall Symp. 1996:155-158.

Spyns P. Natural language processing in medicine: an overview. Meth Inform Med 1996;35:285-301.

Warner AJ, Wenzel PH. A linguistic analysis and categorization of nominal expressions. Proceedings of the 54th Annual Meeting of the American Society for Information Science, Washington, DC, 186-195, 1991.

Figure 1: Sentence diagram for "The man played with the ball."

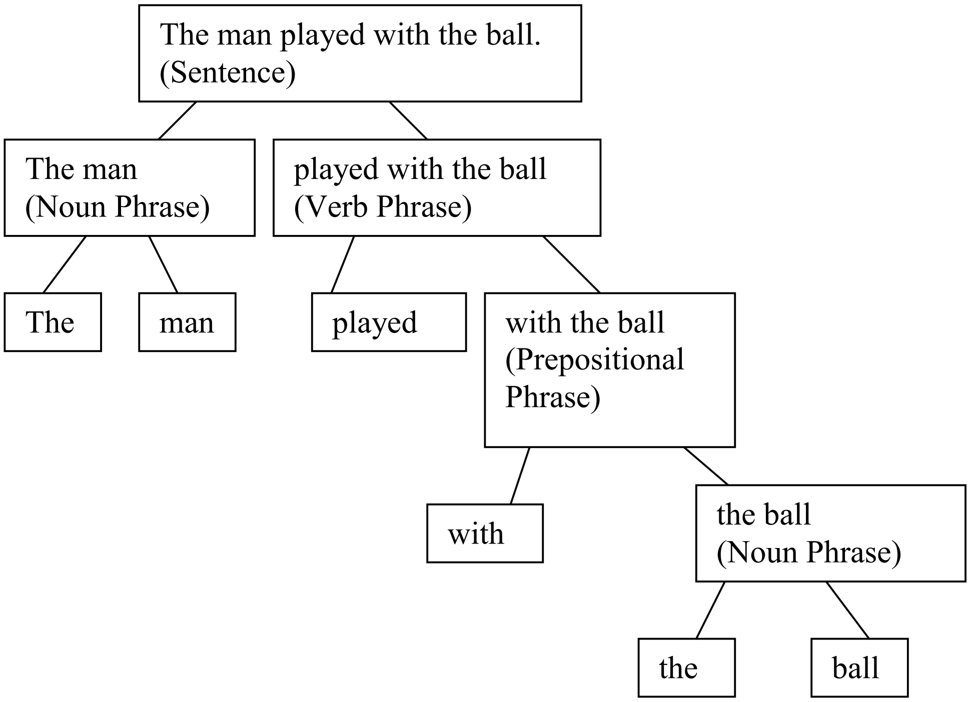
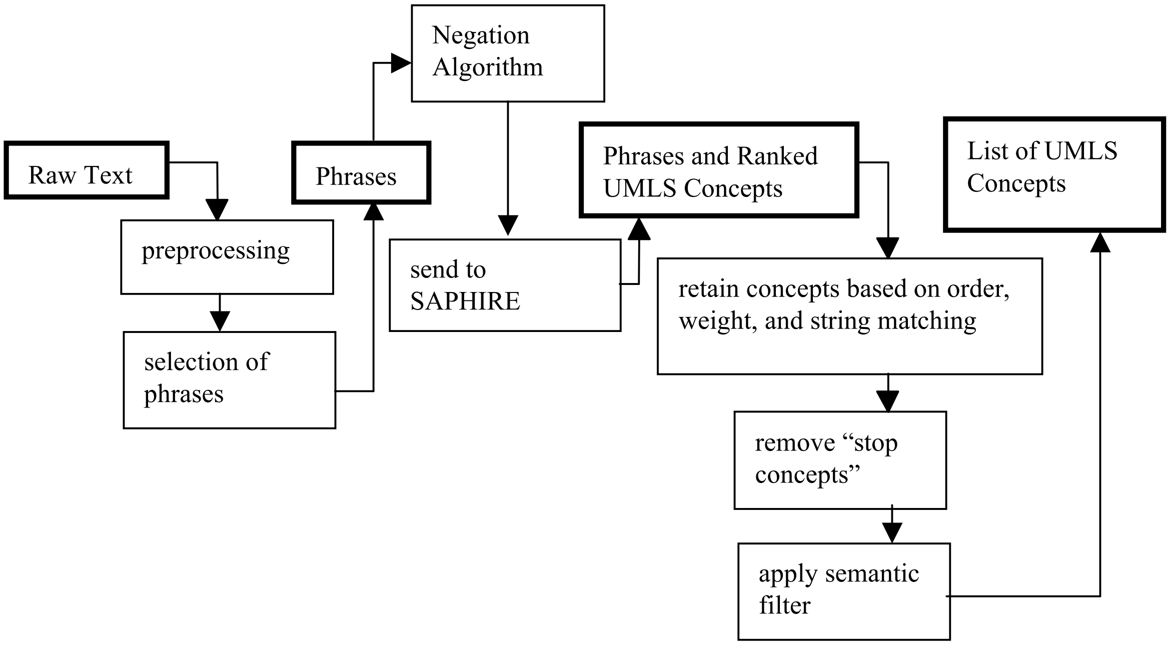

The man played with the ball. (Sentence)
The man (Noun Phrase)
played with the ball (Verb Phrase)
The
man
played
with the ball (Prepositional Phrase)
with
the ball (Noun Phrase)
the
ball

Figure 2: Chart Engine algorithm



Appendices

Appendix A – A sample dictation and indexing concepts

CT_Abd_006.clean

CT SCAN OF THE ABDOMEN WITH LOW OSMOLALITY CONTRAST

HISTORY: METASTATIC RENAL CELL CARCINOMA. NOW PRESENTS WITH ACUTE RIGHT ABDOMINAL PAIN.

Serial unenhanced images of the liver were obtained. After the administration of intravenous contrast, biphasic liver imaging was performed and carried out to the level of the pubic symphysis. Comparison is made to a prior biphasic CT.

Compared with the prior study, there has been increase in size and number of lung base metastatic lesions. A lesion seen abuting the left hemidiaphragm measures 30 cm x 22 cm (series 2, image 6).

Several less than 10 cm hypodense lesions are seen again in the liver and most likely represents cyst. A larger less well defined hypodense lesion is seen in segment 6 of the liver. This measures 20 cm x 20 cm and has also increased in size (series 3, image 48). An additional approximately 10 cm hypodense area is seen adjacent to the falciform ligament and likely represents an area of focal fat, however, an additional metastatic lesion cannot be excluded.

Patient is status post left nephrectomy, and left adrenalectomy. An omental mass that is anterior to the stomach has increased in size from the prior study and is more heterogeneous in appearance. It measures 35 cm x 30 cm (series 3, image 48).

A 10 cm hypodense mass in the right kidney is unchanged from the prior study. Linear filling defect in the distal abdominal aorta is unchanged and consistent with chronic dissection.

The pancreas, gallbladder, spleen and right adrenal gland appear normal. No free air or fluid is seen.

IMPRESSION:

1. INCREASE IN SIZE AND NUMBER OF LUNG BASE METASTASES.
2. INCREASED SIZE IN OMENTAL MASS ANTERIOR TO THE STOMACH.
3. INCREASED SIZE OF HYPODENSE LIVER LESION IN SEGMENT 6 LIKELY REPRESENTING METASTATIC DISEASE. SEVERAL SMALLER HYPODENSE LIVER LESIONS ARE UNCHANGED AND LIKELY REPRESENT CYSTS.
4. UNCHANGED CHRONIC DISTAL AORTIC DISSECTION.
5. UNCHANGED SMALL HYPODENSE RIGHT RENAL PARENCHYMAL MASS.

J30.

My signature below is attestation that I have interpreted this/these examination(s) and agree with the findings as noted above.

END OF IMPRESSION:

Indexing concepts:

C0023884	Liver
C0034015	Pubic Symphysis
C0010709	Cysts
C0230240	Falciform ligament
C0022646	Kidney
C0003484	Aorta, Abdominal
C0012737	Dissection
C0340643	Dissection of aorta
C0220651	Metastasis to lung
C0011980	Diaphragm
C0262613	RENAL MASS
C0202839	Computerized tomography of abdomen
C0278613	metastatic disease
C0028977	Omentum

Appendix B – List of Barrier Words

The words with asterisks were added to the original 248-word list to make the 290-word list:

abnormal*
absent*
about
above
across
after
afterwards
again
against
all
almost
alone
along
already
also
although
always
among
amongst
an
and
another
any
anyhow
anyone
anything
anywhere
appear*
appeared*
appearing*
appears*
are
around
as
at
be
became
because
become
becomes
becoming
been
before
beforehand
behind
being
below
beside
besides
between
beyond
both
but
by
can
cannot
change*
changed*
changes*
changing*
clear*
co
could
demonstrate*
demonstrated*
demonstrates*
demonstrating*
down
during
each
eg
either
else
elsewhere
enough
etc
even
ever
every
everyone
everything
everywhere
except
few
first
for
former
formerly
from
further
had
has
have
he
hence
her
here
hereafter
hereby
herein
herupon
hers
herself
him
himself
his
how
however
ie
if
in
inc
indeed
into
is
it
its
itself
last
latter
latterly
least
less
look*
looked*
looking*

looks*
ltd
many
may
me
meanwhile
measure*
measured*
measures*
measuring*
might
more
moreover
most
mostly
much
must
my
myself
namely
neither
never
nevertheless
next
no
nobody
none
noone
nor
normal*
not
nothing
now
nowhere
of
off
often
on
once
one
only
onto
or
other
others
otherwise
our
ours
ourselves
out
over
own
per
perhaps
present*
rather
represent*
represented*
representing*
represents*
resemble*
resembled*
resembles*
resembling*
same
seem
seemed
seeming
seems
several
she
should
show*
showed*
showing*
shows*
since
so
some
somehow
someone
something
sometime
sometimes
somewhere
still
such
suggest*
suggested*
suggesting*
suggests*
than
that
the
their
them
themselves
then
thence
there
thereafter
thereby
therefore
therein
thereupon
these
they
this
those
though
through
throughout
thru
thus
to
together
too
toward
towards
unclear*
under
until
up
upon
us
very
via
was
we
well
were
what
whatever
when
whence
whenever
where
whereafter

whereas
whereby
wherein
whereupon
wherever
whether
whither
which
while
who
whoever
whole
whom
whose
why
will
with
within
without
would
yet
you
your
yours
yourself
yourselves

Appendix C – Part-of-Speech and Syntactic Tags used by EngLite

EngLite has 16 part-of-speech tags:

Tag	Explanation	Example
A	adjective (or N with a similar core meaning)	blue, sweet
ABBR	Abbreviation	CPR, ACLS
ADV	Adverb	quickly, very
CC	coordinating conjunction	or, and, but
CS	subordinating conjunction	that, which
DET	determiner	the, a, an
EN	non-finite EN form	driven, taken, strayed, played
INFMARK>	infinitive marker (to, in order to)	to, in order to
ING	ING-form	driving, taking, straying, playing
INTERJ	Interjection	Wow!
N	noun (or A or ABBR with a similar core meaning)	boy, dog
NEG-PART	“not”, “n’t”	not, n’t
NUM	Numeral	1
PREP	Preposition	in, by, out, from, to, under
PRON	Pronoun	I, you, me, your, mine, yours, yourself
V	Verb	drive, drives, take, takes, was, is

*First and second columns are taken verbatim from the EngLite documentation.

EngLite has 12 syntactic tags:

Tag	Explanation	Example
&>N	determiner or premodifier of a nominal	*the* boy, *the smart* boy
&NH	nominal head	the *boy*, the smart *boy*, the *boy* in the *yard*
&N<	postmodifier of a nominal	the boy *in* the yard
&>A	premodifying adverb	the *very* smart boy
&AH	adverbial head (for ADV, INTERJ, NEG-PART, PREP)	the boy playing *in* the yard
&A<	Postmodifying adverb (“enough”)	he’s big *enough*

&AUX	Auxiliary	the boy *had* played with the ball
&VP	main verb in a passive verb chain	the ball had been *played* with by the boy
&VA	main verb in an active verb chain	the boy had *played* with the ball
&>CC	introducer of a coordination (“either”, “neither”, “both”)	*both* the boy and the girl, *either* the boy or the girl
&CC	coordinating conjunction	both the boy *and* the girl, either the boy *or* the girl
&CS	subordinating conjunction	the boy *who* played with the ball

*First and second columns are taken verbatim from the EngLite documentation.

Appendix D – Bottom-up Chart-Parsing Algorithm

Allen (1995) describes bottom-up chart-parsing algorithms in Natural Language Processing of sentences. In this work a standard bottom-up chart-parsing algorithm was adapted instead to process EngLite output. Doing so allowed us to form noun phrases using rules that could be easily changed.

This parsing algorithm is *bottom up* because it tries to build bigger pieces from smaller pieces. For example, our algorithm looks at EngLite output and tries to build a
NP4 from a DET and a NP3
and a
NP from an NP4.

The phrase "*chart-parsing*" refers to the fact that the algorithm uses a particular type of data structure called a chart. The chart holds all the information about a sentence, e.g., "word 1 is a noun, word 2 is either a noun or an adjective, words 1 and 2 can form a noun phrase," etc.

The components needed for a bottom-up chart-parser are:
A Chart that stores the completed Constituents
An Agenda that stores the Constituents that still need to be processed.
A list of Active Arcs that stores the Arcs that are not yet completed.

The Arc Extension Algorithm is as follows:

“To add a constituent C from position p1 to p2:

1. Insert C into the chart from position p1 to p2.
2. For any active arc of the form X ← X1 … * C … Xn from position p0 to p1, add a new active arc X ← X1 … C * … Xn from position p0 to p2.
3. For any active arc of the form X ← X1 … Xn * C from position p0 to p1, then add a new constituent of type X from p0 to p2 to the agenda.”

The Chart Parsing Algorithm is as follows:

“Do until there is no input left:

1. If the agenda is empty, look up the interpretations for the next word in the sentence and add them to the agenda.
2. Select a constituent from the agenda (let’s call it constituent C from position p1 to p2).
3. For each rule in the grammar of form X ← C X1 … Xn, add an active arc of form X ← * C X1 … Xn from position p1 to p2.
4. Add C to the chart using the Arc Extension Algorithm above.”

(from Allen, 1995)

Appendix E – Rules for Forming Noun Phrases

The noun phrases in Run 1 and Run 2 conformed to the pattern (regular expression):

{DET} {DET} [ADJ]* [PRE-NOUN]* NOUN-HEAD

as specified in the context-free grammar:

1. NP ← NP1
2. NP ← NP2
3. NP ← NP3
4. NP ← NP4
5. NP1 ← NOUN-HEAD
6. NP2 ← NP1
7. NP2 ← PRE-NOUN NP2
8. NP3 ← NP2
9. NP3 ← ADJ NP3
10. NP4 ← DET NP3
11. NP4 ← DET DET NP3

A DET (determiner) was defined as any word having “DET” (determiner) as one of its parts of speech.

An ADJ (adjective) was defined as any word having syntactic tag &>N (pre-noun modifier) and concomitant part-of-speech not equal to “N” (noun).

A PRE-NOUN (noun functioning as a modifier of a head noun) was defined as any word having syntactic tag &>N (pre-noun modifier) and concomitant part of speech “N.”

A NOUN-HEAD was defined as any word having syntactic tag “&NH.”

The noun phrases in Run 3 conformed to the pattern

NP1 [PREP NP1 [PREP NP1 [PREP NP1 …]]]

where NP1 is a noun phrase as defined above.

All of the PREPs above must have a syntactic tag &<N indicating that the prepositional phrase post-modified a noun head, as opposed to a syntactic tag &AH indicating that the prepositional phrase played an adverbial role.

This was specified by adding the following additional rule:

12. NP ← NP PREP NP