Linguistic Approaches in Information Retrieval of Medical Texts

Anne-Marie Currie, Jocelyn Cohan and Larisa Zlatić

Synthesys Technologies, Inc.

1. Introduction. Medical records contain an enormous amount of specific information about the treatment of individual patients. Much of the specific medical information remains buried in semi-structured electronic text reports, and is difficult for health care providers or medical researchers to find. This paper discusses the linguistic approaches used in retrieving the relevant information from electronic medical records. It is shown how the professional areas of linguistics, information technology and medicine are combined to accurately retrieve the desired information and ultimately improve the care of individual patients.

Information retrieval is achieved through the Clinical Practice Analysis™ (CPA) system developed by Synthesys Technologies, Inc. The linguistic analysis employed in the development of queries within the CPA system enables medical researchers to retrieve information contained in medical records quickly and accurately, and minimizes the amount of irrelevant information retrieved. This is accomplished by translating linguistic generalizations into the CPA search language, called Text Query Language (TQL). The application of core linguistic analyses including syntax, semantics, pragmatics, and sociolinguistics to the query process of medical record texts leads to the identification of clinical issues such as patients at risk for particular conditions, patients eligible for clinical trials and drug interactions. In addition, this collaboration between linguistics, technology and medicine facilitates epidemiological and quality assurance research.

1.1 Description of the CPA System. CPA is a repository of transcribed medical records in which basic document structure characteristics, such as section headings, have been tagged. This minimal tagging process does not require the use of a natural language parser during the initial building of a database. The

linguistic components of CPA include a query engine that can be programmed to identify semantic and syntactic contexts, and a series of "post-build attributes," which identify lexical items in semantic categories, such as family relationships or negation, or syntactic structures associated with particular semantic phenomena, such as negation and modal contexts. Medical records can then be queried using key words in defined linguistic contexts, resulting in a set of documents containing desired information. Pertinent document information can then be extracted from the resulting set into a relational database for analysis by a medical researcher or clinician.

CPA does not employ an automatic knowledge-based natural language processing system. Instead, linguistic generalizations are supplied by the researchers using CPA. This allows for a flexible system that is able to execute searches on millions of documents quickly and effectively. The effectiveness of the system depends in part on the fact that it can be programmed to prevent retrieval of undesirable contexts in which target expressions might occur. For example, in order to ensure that a target expression is in fact linked to the patient and not to another individual, we have developed three strategies. First, we have developed an anaphora attribute that identifies reference to the patient within a document. This attribute allows CPA to link the patient accurately to the particular disease, event, or behavior of interest (cf. (1a)). Second, if appropriate, the family history sections of documents can be excluded from a search because these sections typically do not discuss the medical status and health behaviors of the patient (cf. (1b)). Third, using a post-build attribute that identifies reference to family members and other social relations, we can filter out those instances where family or friends are the referents of a target expression (cf. (1c)).

(1)　　(a) **Patient** was given medications. **He** has a history of **coronary artery disease**.

　　　(b) **Family History**: **Coronary artery disease** in **mother** and **brother**.

　　　(c) **Maternal** history of **coronary artery disease**.

Linguistic analysis of patient medical records has been useful in identifying the difference between desirable and undesirable contexts for target expressions, such as those illustrated in (1a-c). The generalizations derived from this work are incorporated into query strategies. The following sections

discuss specific linguistic observations and how these have been incorporated into an information retrieval system that does not employ natural language processing. Section 2.1 considers relevant issues of word meaning and usage. Section 2.2 addresses the issue of syntactic structure and its relation to both argument structure and lexical-conceptual structure. Section 2.3 considers semantic and pragmatic contexts relevant to the retrieval of desirable documents. Section 3 concludes the paper.

2. Retrieving Relevant Information.

2.1. Lexical Items. Single lexical items or simple phrases can be used to retrieve relevant information. The problem is that often a simple key word search not only returns documents with many irrelevant contexts, it also may not return all the documents with relevant contexts. We have developed dictionaries of related terms that can be easily added to when new contexts are discovered to help address this issue. Three expressions will be used throughout this paper to illustrate specific aspects of linguistic phenomena accessed for accurate information retrieval. Each of these expressions offers distinct properties. In (2a), for example, *infarction* is a specialized medical term that carries different meanings and is not often used in colloquial speech. In contrast, the lexical item *smokes*, in example (2b), is a non-specialized lexical item, used in both the general and medical communities to convey the same unmarked meaning. Finally in (2c), the expression *pronounced* is a lexical item used in both the lay and medical community, but with different meanings.

(2)　　(a)　　History of prior **infarction**.

　　　　(b)　　Patient **smokes**.

　　　　(c)　　He was **pronounced** at 10:00 a.m.

2.1.1. SYNONYMOUS AND NEAR EQUIVALENT EXPRESSIONS. Suppose a medical researcher wants to identify a set of patients who are smokers. Any search engine can run a keyword search like *smok**, which would identify documents containing the morpheme *smok-* in any context. This would, however, fail to return other documents that report a patient is a smoker, through expressions like those in (3a-c):

(3)   (a)   Nicotine use.

        (b) … and nicotine dependence

        (c) … except for tobacco abuse.

Our query to identify smokers needs to take into account expressions such as *"nicotine use"* and *"tobacco abuse."* In addition, we need to filter out those expressions that do not indicate that the patient is a smoker, as in (4). Applying what we know about compounds beginning with *smoke*, the subject of the verb *smoke*, negation, and implicature, we can adequately filter or identify expressions such as those given in (4a-f). A more detailed discussion of the linguistic phenomena is provided in the following sections.

(4)   (a)   concern about smoke inhalation

        (b)   radiation coming from the smoke detector

        (c)   Everyone in the house smokes but they do not smoke around the child.

        (d)   Patient does not smoke.

        (e)   Smoking history is negative

        (f)   Patient must not smoke while on oxygen therapy.

2.1.2. POLYSEMOUS EXPRESSIONS. These are expressions that have multiple meanings for a single lexical item. These distinct meanings of the expressions are related in some way. Within the context of the language used within dictated medical notes, we have identified polysemous expressions that have both marked and unmarked meanings within this specialized context. Knowledge about the multiple meanings of these expressions enables us to develop appropriate queries for topics that call for the use of polysemous expressions. As we develop queries and review data, our knowledge about the source of the data, the expression's location in the document, social habits, and general uses of specific words are considered in the development of any query. An example of such polysemous lexical item is given in (5).

(5)   History of prior **infarction**.

The term *infarction* can be used to indicate several different types of infarctions. The unmarked use in the data we have been working with appears to signify myocardial infarction, while marked uses include cerebral infarctions, lung infarctions, retinal infarctions, and bowel infarctions. However, determining the marked and unmarked uses of an expression depends in part on the source of the documents. For example, a hospital that has expertise in cardiology and serves a higher number of cardiology patients than neurology patients would likely exhibit the unmarked form of *infarction* as myocardial infarction, whereas a hospital that serves a patient community with a high incidence of neurology patients might exhibit the unmarked form of *infarction* as a cerebral infarction. Depending upon the purposes for the query, the source of the data and the needs of the client, the query can be designed to include the unmodified word *infarction*.

The language in medical records also typically reflects the marked and unmarked usage of the broader speech community. Unmarked expressions such as those in (6) report on the habit of smoking tobacco as opposed to other substances.

(6)   (a)   Patient **smokes**.

        (b)   She is a **smoker**.

        (c)   Positive for **smoking**.

The use of such expressions to indicate tobacco smoking is virtually the same in medical records and colloquial speech. The intransitive use of *smoke* in (6a), the agentive nominal in (6b), and gerund form of the verb *smoke* in (6c), all indicate that the patient smokes tobacco, as opposed to marked meanings of smoking marijuana or cocaine.

2.1.3. AMBIGUOUS EXPRESSIONS. Some lexical items identified within medical records are ambiguous. We use this term to describe two separate lexical items with distinct meanings that share the same form. The word *pronounced* is an example of an ambiguous lexical item. One of the meanings of this word occurs in both the lay and medical communities, while another meaning of the word is relegated

to use within the medical field and does not overlap with colloquial usage. Example (7a) illustrates one meaning of *pronounced* that is shared between the general and medical community. Here, *pronounced* is used to indicate that the rash is more significant or severe on the left forearm than the right forearm. This use of *pronounced,* to describe distinctness, appears to be less marked than the use of *pronounced* to report the death of a patient, illustrated by examples (7b-d). The word *pronounced* in examples (7b-e) is a performative used to mean a formal and authoritative announcement of death.

(7)   (a)   The pruritic erythematous rash is more **pronounced** on the left forearm

      than the right.

      (b)   Death was **pronounced**.

      (c)   He was **pronounced** soon thereafter.

      (d)   She was **pronounced**. There was no cardiac activity.

      (e)   The time **pronounced** was 02:25 on 01/25/98.

We incorporate our linguistic knowledge to differentiate the two different senses of *pronounced*. First, these two distinct lexical items differ in their parts of speech. For instance, in example (8a) below, *pronounced$_1$* is used as an adjective to modify the noun phrase *weight loss*. In contrast, example (8b) demonstrates that the lexical item *pronounced$_2$* is a perfomative verb. We argue that the lexeme *pronounced* in example (8b) is the same lexical item used in the expression, "I now **pronounce** you husband and wife" as spoken by someone with the authority to marry two people. In the case of example (8b), the doctor is the person with the authority to perform the official announcement of death and this is most often accomplished with this verb in the passive voice. We speculate that this method of reporting death is one way health providers distance themselves from the death of a patient.

(8)   (a)   The patient admitted to having a **pronounced$_1$** weight loss

      (b)   Death was **pronounced$_2$**.

2.2. Syntactic Contexts. Single lexical items are often not sufficient for precise and efficient information retrieval. In addition to simple words, it is necessary to incorporate phrases in which these lexical items appear. Our text query language uses proximity operators that allow specification of phrase length, thus enabling us to capture various syntactic contexts in which the lexical items may appear. For instance, in all examples in (3a-c) above repeated below in (9), key words *nicotine* and *tobacco*, are the arguments of the nominal heads *use, dependence, abuse*. These arguments precede their nominal heads because they appear in what Grimshaw (1990) calls the synthetic compound constructions.

(9)   (a)   Nicotine use.

        (b) … and nicotine dependence

        (c) … except for tobacco abuse.

In other cases, like in constructions (10a-c) below, the same lexical items, *nicotine* and *tobacco,* are arguments that appear in a complement position, hence, they follow the head noun.

(10)   (a)   use of nicotine

        (b)   dependence on *nicotine*

        (c)   abuse of *tobacco*

What these examples show is that in addition to the linear position of these target words, we need to know their argument structure and how it maps to syntactic structure. The observed linguistic patterns of argument linking are incorporated into queries for efficient and accurate information retrieval.

Valence alternations of certain lexical items should also be incorporated in information retrieval techniques. To illustrate, the verb *smoke* can be used both transitively and intransitively, as shown by the following examples.

(11)   Intransitive uses of *smoke*:

        (a)   He doesn't *smoke.*

(b) He has never *smoked.*

(c) The patient does *smoke*.

(d) She also *smokes* although she is working on quitting at the present time

(12) Transitive uses of *smoke*:

  (a) He no longer *smokes* cigarettes.

  (b) He used to *smoke* cigars.

  (c) He currently *smokes* three cigarettes per day.

  (d) The patient *smokes* occasional cigarettes, two to three per day.

  (e) The patient *smokes* one pack per day

The two different verb types seem to correspond to different interpretations. Specifically, the intransitive forms of *smoke* shown in (11), are used in contexts in which the habit of smoking or not smoking is being described. The transitive forms are used for all other purposes, such as negative contexts where the emphasis is on non smoking (as in (12a), or for describing smoking of something other than the cigarettes, as in (12b). However, the most frequent use of the transitive form of *smoke* is found for describing the amount of cigarettes the patient consumes, as the examples in (12c-e) illustrate. This observed correlation between the verb's valence and the semantic context facilitates the process of information retrieval.

A similar valency alternation is found with the verb *pronounce* when used to state that a patient has died. This verb generally takes a secondary predicate phrase, as shown in (13). In these examples, the predicative adjective phrases, *dead*, *deceased,* and *expired* function as a complement of the passive verb *pronounced.*

(13) (a) The patient was *pronounced dead* by Dr. Smith at 19:24.

  (b) The patient was *pronounced dead* at 4:30 on the date of admission.

  (c) The patient was *pronounced deceased* at 10:50 a.m. on March 29, 1999.

  (d) The patient was *pronounced expired* at 7:45 p.m. on 06/08/95.

However, medical reports quite often contain the examples in which the complement adjective phrase is completely omitted, as shown in (14).

(14)   (a)   After the patient had been *pronounced*, he has been moved to the morgue.

           (b)   Death was *pronounced.*

           (c)   He was *pronounced* soon thereafter.

           (d)   The patient was *pronounced* at 10:55 a.m., and his family was notified.

           (e)   The efforts were terminated at 0115 and patient *pronounced* at that time.

           (f)   The patient was *pronounced* in the Emergency Department and autopsy

                was requested.

The first three examples in (14) show that the verb *pronounced* does not require any complements at all, although the complement prepositional phrases, indicating place and time, are generally present, as in (14d-f). These examples thus show that semantic arguments are not always overtly expressed in syntax. It is either through the pragmatic context or through a single, 'salient' argument that the other semantic arguments of the verb are recovered. For example, in (14c), the precise meaning of the passive verb *pronounced* is induced by the presence of the anaphoric pronominal subject *he* that refers to the patient. In (14b), the subject noun phrase, *death,* acts as a salient argument that determines the specific meaning of the verb *pronounced*.

       As mentioned earlier, the passive verb is more often used for stating the patient's death than the active verb form. However, regardless of which verb form of *pronounced* is used, we find the same argument alternations, as shown in (15).

(15)   (a)   I *pronounced the patient dead* at 1915.

           (b)   I *pronounced the patient deceased* at approximately 0727 hours on

                September 23, 1999.

           (c)   I *pronounced the patient,* with his family at the bedside, at 8:04 a.m.

(d)   His family was present as *I pronounced him*, and funeral arrangements will be in the

paper tomorrow.

These examples of argument alternations have not only theoretical linguistic significance, but also significance for information retrieval techniques. For linguistic theory, these facts further confirm that the 'syntactic' argument structure must be distinguished from the semantic structure, or what Grimshaw (1990) calls the lexical conceptual structure. In order to meet information retrieval needs, computational methods are devised to account for argument alternation and reconstruction of implicit arguments.

2.3. Semantic and Pragmatic Issues. At the root of information retrieval is the issue of semantics. Retrieval on the basis of key expressions is effective only insofar as these expressions carry the desired meaning. Section 2.1 addressed the role of word meaning; section 2.2 addressed the role that syntax plays in determining the meaning of a key search term. The larger context in which a target expression occurs can also be crucial to its meaning. In this section we consider some sentence- and discourse-level semantic and pragmatic issues.

There are numerous contexts in which target expressions may occur and yet may not be connected to a report about the patient. We previously saw examples of one such context, where target expressions are linked to someone other than the patient (cf. (1b-c)). Two additional contexts that must also be considered in the retrieval of information from medical records are belief contexts and downward entailing contexts, both long recognized in semantics literature (e.g., Quine 1961). Also to be considered are the roles that presupposition and implicatures may play in signaling desirable and undesirable documents.

2.3.1. BELIEF CONTEXTS. Belief contexts are introduced by words like *believe, suspect* or *think*. Propositions embedded in such contexts need not be true. Occurrence of a target term in a document in a belief context is thus not really any more informative than if the target term had not occurred. That is, if a researcher is looking for documents that report a definitive diagnosis of myocardial infarction or heart attack, documents containing excerpts like those in (16) would be desirable only if the diagnosis were confirmed *elsewhere* in the document.

(15)   (a)   He apparently had chest pain and it *was thought* he was having a **heart attack**

         (b)   It was *believed* that the patient had a non-Q wave **myocardial infarction**.

Neither the proposition *he was having a heart attack* in (16a), nor the proposition *the patient had a non-Q-wave myocardial infarction* in (16b) need be true (although either might be). While these excerpts might be considered informative in that they reveal that the medical staff *considered* a diagnosis of myocardial infarction, without additional confirmation they cannot be understood to convey that this diagnosis was actually made. Thus, these are not desirable occurrences of the target terms if our goal is to identify a patient who had a confirmed heart attack.

Some diagnoses or behaviors occur more often than others in belief contexts. For example, while it is fairly common to find *myocardial infarction* (and its synonyms) in excerpts like those in (16) we do not typically find excerpts like *it was thought that the patient was a smoker.* Thus, the importance of taking belief-contexts into account in the retrieval of documents varies depending on the goals of a project. The flexibility of the CPA system allows researchers to incorporate belief contexts into queries when necessary for the goals of a particular project.

2.3.2. DOWNWARD-ENTAILING CONTEXTS. Other circumstances in which propositions do not necessarily hold occur in so-called downward entailing contexts (Ladusaw 1980). Two such contexts that commonly appear in medical records are negative and conditional contexts, as seen in (17) and (18), respectively.

(16)   (a) He ***denied*** any alcohol use or ***any*** **tobacco use**

         (b) He is ***not*** aware that he has ***ever*** had a **heart attack** or ***any*** significant **cardiac problems** in
              the past.

         (c) He also had cardiac enzymes drawn that did ***not*** show ***any*** evidence of **infarction.**

         (d) It is ***not*** associated with **chest pain** or shortness of breath.

In the context introduced by *denied* in (17a), the concept *tobacco use* includes chewing tobacco and smoking cigarettes, cigars or pipes. Thus, if it is true that the patient does not use tobacco, then it is also true that he does not chew or smoke tobacco. The proposition that the patient does not use tobacco <u>entails</u> that he does not chew or smoke it. In ordinary use, *deny* can convey the idea that the speaker or writer does not believe the denial, which is not the case when *deny* is used in medical records.

Some examples of conditional contexts are shown below.

(17)   (a) ***If*** she develops more fevers, chills, **chest pain** or sputum changes color, she will call me immediately.

   (b) I want to see ***if*** there is ***any*** change in his EKG to be sure that he has ***not*** had a silent **infarction.**

   (c) The patient had some difficulty with dementia and it is not clear ***whether*** he is actually experiencing some **chest pain** but I think rather not.

In the context introduced by *if* in (18a), the concept *chest pain* includes chest pain due to injury, respiratory infection and cardiac problems. Thus, the proposition that the patient will call if she develops chest pain <u>entails</u> that she will call if she develops chest pain due to respiratory infection, injury or other causes.

Recognition of these contexts is important to the accurate retrieval of relevant records. The excerpts in (17) and (18) do not assert that the patient uses tobacco, has had a heart attack, or has experienced chest pain, and so would not be desirable documents if a researcher is looking for positive occurrences of these conditions. Target terms in downward-entailing contexts like these are typically no more informative than target terms in belief contexts.

Certain predictable lexical items typically introduce negative and conditional contexts, *deny, not, never, no, if, whether,* etc. The appearance of other lexical items – negative polarity items (NPIs) like *any* and *ever,* is also typically limited to such contexts. We have developed and made use of post-attributes that specifically identify lexical items associated with such contexts. These post-attributes can be used to exclude documents that contain search terms only in these contexts and can thus be used to exclude

undesirable documents. Again, these undesirable documents contain the target terms, but do not in fact assert that the target term applies to the patient.

In medical records, as in other kinds of discourse, we find downward-entailing contexts that are not marked by the explicit appearance of these already-identified lexical items. For example, contexts like imperatives, are frequently used as conditionals in medical records. In (19), for example, the clause *return for chest pain* represents instructions to the patient to return if she experiences chest pain.

(18)   DISCHARGE *INSTRUCTIONS*:...3. Return for **chest pain**, other weakness, dizziness …

If a researcher were searching for documents that report on patients with chest pain, this would not necessarily be a desirable document. Because the imperative form is ambiguous with other verb forms, identifying the imperative solely by its morphological form is not a viable search strategy. There are also no NPIs or other lexical items like those associated with negative and conditional contexts to help identify the downward entailing context. The excerpt in (19) shows, however, that the occurrence of the search term falls within *instructions* to the patient. Additional lexical items, like *instructions,* that are typically associated with downward-entailing contexts in medical records, can be incorporated into our strategy for identifying such contexts, preventing potentially undesirable documents from being retrieved.

2.3.3. PRESUPPOSITION. Belief and downward entailing contexts sometimes contain truly desirable targets. This occurs when the target is connected to a presupposition. Presuppositions typically "survive" belief and downward entailing contexts (Beaver 1997). Presuppositions can be assumed to be true – in fact, typically must be assumed to be true – even when the propositions containing them are not.

The definite noun phrases *his myocardial infarction* in (20a) and *the earlier episode of chest pain* in (20b) have existential presuppositions that survive the negative and belief contexts in which they occur. The example in (20a) presupposes that the patient had a myocardial infarction and the example in (20b) that the patient had an episode of chest pain.   Additionally, adjectives like *earlier* in (20b) and *recurrent* (20c) also bring existential presuppositions to the noun phrases in which they occur. Thus, even indefinite noun phrases, like *recurrent breast cancer* in (20c), can carry an existential presupposition. Note that while the

recurrence of breast cancer does not survive the negative context, the presupposition that the patient had breast cancer at some time in the past does.

(19)   (a)   The patient remained stable … having **no** acute complications from ***his* myocardial infarction** and thrombolytic therapy.

presupposition: the patient had a myocardial infarction.

(b)   Patient **believed** that ***the earlier episode of* chest pain** had only lasted two to three seconds.

presupposition: The patient had an episode of chest pain at some time in the past.

(c)   **No** evidence of ***recurrent* breast cancer** five years postoperatively.

presupposition: the patient had breast cancer at some time in the past.

Verbal elements can also carry presuppositions.   In (21a), *quits smoking* conveys the presupposition that the patient (referred to by the reporting physician here as *he*) is a smoker, which survives both conditional and belief contexts. In (21b), *recurs* conveys the presupposition that the patient had chest pain, which here survives a conditional context (although, again, the *recurrence* does not).

(20)   (a) I ***believe*** that ***if*** he ***quits* smoking** now, he will do fine in the future.

presupposition: the patient is a smoker.

(b) ***If* chest pain *recurs***, we will consider referring her for additional testing.

presupposition: the patient had chest pain.

Lexical items that contribute presuppositions can be incorporated into queries so that potentially desirable excerpts like those in (20) and (21) are not excluded by efforts to avoid returning documents containing target terms in downward-entailing or belief contexts.

2.3.4. IMPLICATURE. In many cases, documents about patients do not presuppose or directly state that a condition or behavior is part of a patient's medical history. Many documents simply imply these conclusions. Implicatures differ from presuppositions because they can be cancelled in many situations by the addition of further information.

Pragmatic knowledge makes an implicature conveyed in medical records very strong. For example, from (22a), a reader can be fairly certain that the patient is a smoker because we know that typically, nicotine patches are not worn by people who don't smoke. From (22b), a reader can be fairly certain that the patient has a history of asthma attacks, because a physician would be unlikely to mention that a patient had not had an attack since her last visit if she had never had one. Nevertheless, these conclusions are implicatures, because they could theoretically be cancelled. Perhaps the patient in (22a) has been prescribed a nicotine patch because he chews tobacco. Perhaps the patient in (22b) was expected to develop an asthma condition for some reason – say, if her last visit to the clinic was precipitated by exposure to a chemical that can trigger development of asthma in some victims – but has still never actually had an asthma attack.

(21) (a) Special discharge instructions include absolutely ***no* smoking** while ***wearing nicotine patch***.

implicature: The patient is a smoker.

(b) She has ***not*** had an **asthma** attack since ***her last visit to the clinic***.

implicature: The patient has a history of asthma attacks.

The conditions necessary to cancel the implicatures in (22) are rather unlikely to occur. The excerpts in (22) can, for all intents and purposes, be considered to represent desirable occurrences of the target terms, since they carry an implicature that the target term applies to the patient, and this implicature is not likely to be cancelled.

In other cases, implicatures are less strong, and can be much more easily cancelled. Excerpts containing these weaker implicatures are typically ambiguous. More information will always be required in order to decide whether the patient matches the desired clinical criteria or not.

(22)   (a)   Special discharge instructions include absolutely *no* **smoking**.

              implicature: The patient is a smoker.

        (b)   The patient is a 51-year-old male with *no previous* **heart disease**, who developed chest pain

              the day before his transfer.

              implicature: The patient has now been diagnosed with heart disease.

Medical staff may warn a patient not to smoke even if the patient were not a habitual smoker. This might

occur if the patient were being treated for a condition like asthma or treated with bottled oxygen.   The

implicature of (23a) would not necessarily remain in this context. Medical staff may consider a diagnosis

of heart disease in patients who are at risk for it because of age, gender or other factors, and later rule this

possibility out. The implicature of (23b) would be cancelled by a report that the patient's chest pain on the

occasion in question was caused by something other than heart disease.

    We can identify documents containing implicatures that can potentially be cancelled with specific

queries that capture the relevant contexts.   The relevant subset of documents or contexts can be isolated

and provided to client-researchers for their review so that these can be categorized according to the goals of

the research.

    3. Conclusion. In this paper we provided a linguistic analysis of three medical topics: infarctions,

tobacco use, and the pronouncement of death, in order to illustrate common issues involved in information

retrieval from electronic medical records. In general, all topics will have similar linguistic issues associated

with the retrieval of desired information and the exclusion of undesired information regarding patient

conditions. Specifically, by applying the linguistic generalizations discussed in this paper and additional

analyses, we are able to identify a population of patients who have a clinical profile of having had a heart

attack, continue to abuse tobacco and thus, are potentially at risk for experiencing another heart attack. In

addition, we can identify a sub-population of this group who have been advised to quite smoking and who

are non-compliant with this advice. The identification of patients who have died will enable a researcher or

physician to eliminate patients who are not eligible for a clinical trial and eliminate the potential mistake of

mailing an invitation of trial participation to family members of the deceased.

The application of linguistic methods to the retrieval of medical information, can result in improving the quality of patient care and the efficiency of the institution responsible for the administration of patient care.   Through the discussions presented in this paper, we have demonstrated how the professional areas of linguistics, technology and medicine are coordinated to retrieve information that has the potential to impact large organizations, such as hospitals, as well as affect change at an individual level, namely, the patient.

The Clinical Practice Analysis system enables the medical researcher to create powerful queries to retrieve information about individual patients or entire patient populations quickly and easily, allowing researchers in health care to make use of the vast quantity of data available in medical records.   The efficiency of this system has been enhanced by the implementation of linguistic generalizations specific to medical records, such as lexical variation, ambiguity, argument alternation, belief contexts, downward-entailing contexts and presupposition. This paper has illustrated that an interdisciplinary team of linguists, programmers and clinical professionals can impact patient's health at both the individual and institutional levels.

References

Beaver, (1997). Presupposition.   In J. van Bentham and A. ter Meulen, eds, *Handbook of Logic and Language.* Elsevier, Amsterdam.

Grimshaw, J. (1990). *Argument Structure*. The MIT Press, Cambridge, Mass.

Ladusaw, W. (1980). *Polarity Sensitivity as Inherent Scope Relations*. Indiana University Linguistics Club.

Quine, W. (1961). *From a Logical Point of View.* The MIT Press, Cambridge, Mass.