8 Natural Language Processing in Health Care and Biomedicine

Carol Friedman and Noémie Elhadad

After reading this chapter, you should know the answers to these questions:

- Why is natural language processing important?
- What are the potential uses for natural language processing (NLP) in the biomedical and health domains?
- What forms of knowledge are used in NLP?
- What are the principal techniques of NLP?
- What are challenges for NLP in the clinical, biological, and health consumer domains?

8.1 Motivation for Natural Language Processing

Natural language is the primary means of human communication. In biomedical and health areas,[1] knowledge and data are disseminated in textual form as articles in the scientific literature, as technical and administrative reports on the Web, and as textual fields databases. In health care facilities, patient information mainly occurs in narrative notes and reports. Because of the growing adoption of electronic health records and the promise of health information exchange, it is common for a patient to have records at multiple facilities, and for a chart of a single patient at one institution to comprise several hundred notes. Because of the explosion of online textual information available, it is difficult for scientists and health care professionals to keep up with the latest discoveries, and they need help to find, manage, and analyze the enormous amounts of online knowledge and data. On the Web, individuals exchange and look for health-related information, and health consumers and patients are often overwhelmed by the amount of the information available to them, whether in traditional websites or through online health communities. There is also much information disseminated verbally through scientific interactions in conferences, in care teams at hospitals, and in patient-doctor encounters. In this chapter however, we focus on the written form.

While there is valuable information conveyed in text, it is not in a format amenable to further computer processing. Texts are difficult to process reliably because of the inherent characteristics and variability of language. Since structured standardized data are more useful for most automated applications, a significant amount of manual work is currently devoted to mapping textual information into a structured or coded representation: in the clinical realm, for instance, professional

[1] Unless stated otherwise, the general domain and the topics of text materials discussed in this chapter refer to biomedicine and health.

C. Friedman, PhD (✉) • N. Elhadad, PhD
Department of Biomedical Informatics,
Columbia University, 622 West 168th Street,
VC Bldg 5, New York 10032, NY, USA
e-mail: friedman@dbmi.columbia.edu;
noemie@dbmi.columbia.edu

This chapter is adapted from an earlier version in the third edition authored by Carol Friedman and Stephen B. Johnson.

E.H. Shortliffe, J.J. Cimino (eds.), *Biomedical Informatics*,
DOI 10.1007/978-1-4471-4474-8_8, © Springer-Verlag London 2014

coders assign billing codes corresponding to diagnoses and procedures to hospital admissions; indexers at the National Library of Medicine assign MeSH terms to represent the main topics of scientific articles; and database curators extract genomic and phenotypic information on organisms from the literature. Because of the overwhelmingly large amount of textual information, manual work is very costly, time-consuming, and impossible to keep up to date. One aim of **natural language processing** (NLP) is to facilitate these tasks by enabling use of automated methods that represent the relevant information in the text with high validity and reliability.

Another aim of NLP is to help advance many of the fundamental aims of biomedical informatics, which include the discovery and validation of scientific knowledge, improvement in the quality and cost of health care, and support to patients and health consumers.

The massive amounts of texts amassed through clinical care or published in the scientific literature or on the Web can be leveraged to acquire and organize knowledge from the information conveyed in text, and to promote discovery of new phenomena. For instance, the information in patient notes, while not originally entered for discovery purposes, but rather for the care of individual patients, can be processed, aggregated and mined to discover patterns across patients. This process, commonly referred to as secondary use of data, shows much promise for some of the current challenges of informatics, such as **comparative effectiveness research**, **phenotype definition**, **hypothesis generation** for clinical research and understanding of disease, and **pharmacovigilance**. For the literature, NLP can speed up the high throughput access needed for scientific discoveries and their meta-analysis across individual articles, by identifying similar results across articles (i.e. recent treatments for diseases, reports of adverse drug events), and by connecting pieces of information among articles (i.e. constructing biomolecular pathways).

For clinicians interacting with an electronic health record and treating a particular patient, NLP can support several points in a clinician workflow: when reviewing the patient chart, NLP can be leveraged to aggregate and consolidate information spread across many notes and reports, and to highlight relevant facts about the patient. During the decision-making and actual care phase, information extracted through NLP from the notes can contribute to the decision support systems in the EHR. Finally, when health care professionals are documenting patient information, higher quality notes can be generated with the help of NLP-based methods.

For quality and administrative purposes, NLP can signal potential errors, conflicting information, or missing documentation in the chart. For public health administrators, EHR patient information can be monitored for **syndromic surveillance** through the analysis of ambulatory notes or chief complaints in the emergency room.

Finally, NLP can support health consumers and patients looking for information about a particular disease or treatment, through automated **question understanding** which can then facilitate better access to relevant information, targeted to their information needs, and to their health literacy levels through the analysis of the topics conveyed in a document as well as the vocabulary used in the document.

Across all these use cases of NLP, the techniques of natural language processing provide the means to bridge the gap between unstructured text and data by transforming the text to data in a computable format, allowing humans to interact using familiar natural language, while enabling computer applications to process data effectively and to provide users with easy access and synthesis of the raw textual information. The next section gives a more in-depth definition of NLP and a quick history of NLP in biomedicine and health. Section 8.3 presents some of the well-studied applications of NLP. Sections 8.4 and 8.5 introduce linguistic background and ways in which linguistic knowledge is leveraged to build NLP tools. Section 8.5 also focuses on evaluation methodology for NLP-based systems. Section 8.6 provides a discussion of the challenges entailed in processing texts in the biological, clinical and general health domains, which are currently active areas of research in the NLP community. Finally Sect. 8.7 provides pointers to useful resources for NLP research and development.

8.2 What Is NLP

Natural language processing is currently a very active and exciting research area in informatics. The term NLP is often used to include a group of methods that involve the processing of unstructured text, although the methods themselves range widely in their use of knowledge of language. Some methods use minimal linguistic knowledge, and are based solely on the presence of words in text. For these methods, the only linguistic knowledge needed is the knowledge of what constitutes a word; these methods often depend on a **keyword** or **bag-of-words** approach. One example of a method that uses only words is a search engine that retrieves documents containing the presence of a combination of words in a collection, although these words may have no relation to each other in the retrieved documents. Another example is a **machine learning** method that uses words independently of each other as features when building a statistical model. Other NLP methods contain more advanced knowledge of language, and these are the methods that are the focus of this chapter. Generally, these more advanced linguistic methods aim to determine some or all of the syntactic or semantic structure in text and to interpret some of the meaning of relevant information in text.

The reason why it is possible to process natural language using computational methods is that language is formulaic: it consists of discrete symbols (words), and rules (a grammar) specifying how different linguistic elements can be combined to create a sequence of words that represents a well-formed sentence or phrase that conveys a particular meaning. According to Harris (1982), it is possible to represent the content of a sentence in an operator argument structure, similar to formulations in **predicate logic**. The formulaic aspect of language can also explain why machine-learning approaches have been successful at some NLP tasks. In particular, patterns, when present in large amounts of text can be detected automatically.

Early work in NLP began in the 1950s. In the 1960s and 1970s, some successful general language NLP systems were developed that involved a very limited domain along with highly specific tasks, such as Eliza (Weizenbaum 1966), SHRDLU (Winograd 1972), and LUNAR (Woods 1973). In the 1970s, the Linguistic String Project (LSP) under the leadership of Dr. Naomi Sager, a pioneer in NLP, developed a comprehensive computer grammar and parser of English (Grishman et al. 1973; Sager 1981), and also began work in NLP of clinical reports (Sager 1972, 1978; Sager et al. 1987) that continued into the 1990s. A number of other clinical NLP systems were developed starting in the late 1980s and early 1990s, and are discussed in more detail by Spyns (Spyns 1996). Some clinical NLP systems that were associated with numerous publications in that period include SPRUS (which evolved into Symtext and then MPLUS) (Haug et al. 1990, 1994; Christensen et al. 2002), MedLEE (Friedman et al. 1994; Hripcsak et al. 1995), the Geneva System (Baud et al. 1992, 1998), MeneLAS (Zweigenbaum and Courtois 1998), and MEDSYNDIKATE (Hahn et al. 2002). NLP processing of the literature started to take hold in the late 1990s with the large increase in publications concerning biomolecular discoveries and the need to facilitate access to the information. Early work in the biomolecular NLP area involved identification of biomolecular entities in text (Fukuda et al. 1998; Jenssen and Vinterbo 2000), and then extraction of certain relations between the entities (Sekimizu et al. 1998; Rindflesch et al. 2000; Humphreys et al. 2000; Park et al. 2001; Yakushiji et al. 2001; Friedman et al. 2001). Meanwhile, in the clinical literature, similar work was carried out to recognize mentions of disorders, findings, and treatments (Aronson 2001) as well as certain relationships, such as diseases and their corresponding treatments in order to mine results conveyed across clinical studies (Srinivasan and Rindflesch 2002).

The assumption concerning much of the early NLP work was that successful NLP systems required substantial knowledge of language integrated with real-world knowledge in order to process text and solve real-world problems. Thus, the primary focus of much of the early work concerned representation of syntactic and semantic knowledge, which was a complex task generally requiring linguistic expertise (Sager 1981; Grishman and Kittredge 1986) along with development of rule-based systems that used the knowledge to parse and interpret the text. The trend started to shift in the biomedical domain from rule-based systems to probabilistic NLP systems in the early 2000s, likely due to the availability of large collections of annotated textual material in the general language domain (Marcus et al. 1993; Palmer et al. 2005) and in the biomolecular domain (Kim et al. 2003), to the rise of machine learning approaches that use the text collections to uncover patterns in text (Manning and Schütze 1999; Bishop 2007).

NLP is multi-disciplinary at its core, weaving together theories of linguistics, computation, representation, and knowledge of biology, medicine and health. Within the computational field itself, NLP intersects with many fields of research, including computational linguistics, knowledge representation and reasoning, knowledge and information management, and machine and statistical learning. Furthermore, because NLP tools are often part of systems targeted at end-users, NLP also intersects with the field of human-computer interaction and cognitive science. The inter-disciplinary nature of NLP has important implications for the design, development and evaluation of NLP-based systems. For instance, it would be impossible to perform a proper error analysis of an NLP tool without expertise in the domain, whether biological or clinical.

8.3 Applications of NLP

Natural language processing has a wide range of potential applications. The following are important applications of NLP technology for biomedicine and health:

- **Information extraction**, the most common application of NLP in biomedicine, locates and structures specific information in text, usually without performing a complete linguistic analysis of the text, but rather by looking for patterns in the text. Once textual information is extracted and structured it can be used for a number of different tasks. In **biosurveillance**, for instance, one can extract symptoms from a chief complaint field in a note written for a patient admitted to the emergency department of a hospital (Chapman et al. 2004) or from ambulatory electronic health records (Hripcsak et al. 2009). The extracted data, when collected across many patients, can help understand the prevalence as well as the progression of a particular epidemic. In biology, biomolecular interactions extracted from one article or from different articles can be merged to construct biomolecular pathways. Figure 8.1 shows a pathway in the form of a graph that was created by extracting interactions from one article published in the journal Cell (Maroto et al. 1997). In the clinical domain, pharmacovigilance systems can use structured data obtained by means of NLP on huge numbers of patient records to discover adverse drug events (Wang et al. 2009a).
- The techniques for information extraction may be limited to the identification of names of people or places, dates, and numerical expressions, or to certain types of terms in text (e.g.

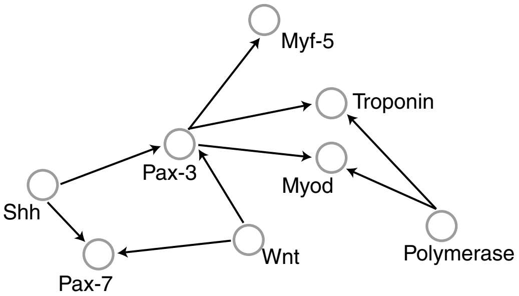


Fig. 8.1 A graph showing interactions that were extracted from an article. A vertex represents a gene or protein, and an edge represents the interaction. The *arrow* represents the direction of the interaction so that the agent is represented by the outgoing end of the arrow and the target by the incoming end

mentions of medications or proteins), which can then be mapped to canonical or standardized forms. This is referred as **named-entity recognition** or **named-entity normalization**. More sophisticated techniques identify and represent the modifiers attached to a named entity. Such advanced methods are necessary for reliable retrieval of information because the correct interpretation of a term typically depends on its relation with other terms in a given sentence. For example, the term *fever* has different interpretations in *no fever*, *high fever*, *fever lasted 2 days*, and *check for fever*. Defining the types of **modifiers of interest** (e.g. *no* is a negation modifier, while *lasted 2 days* is a temporal modifier), as well as techniques to recognize them in text, is an active topic of research. Identifying **relations among named entities** is another important information extraction method. For example, when extracting adverse events associated with a medication, the sentences "*the patient developed a rash from amoxicillin*" and "*the patient came in with a rash and was given benadryl*" must be distinguished. In both sentences, there is a relation between a rash and a drug, but the first sentence conveys a potential adverse drug event whereas the second sentence conveys a treatment for an adverse event. As entities are extracted within one document or across documents, one important step consists of **reference resolution**, that is, recognizing that two mentions in two different textual locations refer to the same entity. In some cases, resolving the references is very challenging. For instance, mentions of *stroke* in two different notes associated with the same patient could refer to the same stroke or two different strokes; additional contextual information and domain knowledge is often needed to resolve this problem.

- **Information retrieval** (IR) and NLP overlap in some of the methods that are used. IR is discussed in Chap. 21, but here we discuss basic differences between IR and NLP. IR methods are generally geared to help users access documents in large collections, such as electronic health records, the scientific literature, or the Web in general . This is a crucial application in biomedicine and health, due to the explosion of information available in electronic form. The essential goal of information retrieval is to match a user's query against a document collection and return a ranked list of relevant documents. A search is performed on an **index** of the document collection. The most basic form of indexing isolates simple words and terms, and therefore, uses minimal linguistic knowledge. More advanced approaches use NLP-based methods similar to those employed in information extraction, identifying complex named entities and determining their relationships in order to improve the accuracy of retrieval. For instance, one could search for *hypertension* and have the search operate at the concept level, returning documents that mention the phrase *high blood pressure* in addition to the ones mentioning *hypertension* only. In addition, one could search for *hypertension* in a specific context, such as treatment or etiology.
- **Question answering** (QA) involves a process whereby a user submits a natural language question that is then automatically answered by a QA system. The availability of information in journal articles and on the Web makes this type of application increasingly important as health care consumers, health care professionals, and biomedical researchers frequently search the Web to obtain information about a disease, a medication, or a medical procedure. A QA system can be very useful for obtaining the answers to factual questions, like "*In children with an acute febrile illness, what is the efficacy of single-medication therapy with acetaminophen or ibuprofen in reducing fever*?" (Demner-Fushman and Lin 2007). QA systems provide additional functionalities to an IR system. In an IR system, the user has to translate a question into a list of keywords and generate a query, but this step is carried out automatically by a QA system. Furthermore, a QA system presents the user with an actual answer (often one or several passages extracted from the source documents), rather than a list of relevant source documents. QA

has focused on the literature thus far (Demner-Fushman and Lin 2007; Cao et al. 2011).

- **Text summarization** takes one or several documents as input and produces a single, coherent text that synthesizes the main points of the input documents. Summarization helps users make sense of a large amount of data, by identifying and presenting the salient points in texts automatically. Summarization can be generic or query-focused (i.e. taking a particular information need into account when selecting important content of input documents). Query-focused summarization can be viewed as a post-processing of IR and QA: the relevant passages corresponding to an input question are further processed into a single, coherent text. Several steps are involved in the summarization process: content selection (identifying salient pieces of information in the input document(s)), content organization (identifying redundancy and contradictions among the selected pieces of information, and ordering them so the resulting summary is coherent), and content re-generation (producing natural language from the organized pieces of information). Like question answering, text summarization has also focused on the literature (Elhadad et al. 2005; Zhang et al. 2011).
- Other tasks: **Text generation** formulates natural language sentences from a given source of information that is not directly readable by humans. Generation can be used to create a text from a structured database, such as summarizing trends and patterns in laboratory data (Jordan et al. 2001). **Machine translation** converts text in one language (e.g. English) into another (e.g. Spanish). These applications are important in multilingual environments in which human translation is too expensive or time consuming (Deleger et al. 2009). **Text readability assessment and simplification** is becoming relevant to the health domain, as patients and health consumers access more and more medical information on the Web, but need support because their health literacy levels do not match the ones of the documents they read (Elhadad 2006; Keselman et al. 2007). Finally, **sentiment analysis and emotion detection** are slightly more recent applications of NLP and belong to the general task of automated content analysis. There are promising research results showing that patients' discourse can be analyzed automatically to help detect their mental states (Pestian and Matykiewicz 2008).

8.4 Linguistic Levels of Knowledge and Their Representations

While current linguistic theories differ in certain details, there is broad consensus that linguistic knowledge consists of multiple levels: **morphology** (words and meaningful parts of words), **syntax** (structure of phrases and sentences), **semantics** (meaning of words, phrases, and sentences), **pragmatics** (impact of context and of intent of the speaker on meaning), and **discourse** (paragraphs and documents). Human language processing may appear deceptively simple, because we are not conscious of the effort involved in learning and using language. However, a long process of acculturation is necessary to attain proficiency in speaking, reading, writing, and understanding with further intensive study to master a different language or the specialized languages of biological science and medicine. This section introduces the types of knowledge entailed in each of the levels as well as their representations. Traditionally, lexicography (the study of the morphology, syntax and semantics of words), the development of rules, or grammars, and the acquisition of linguistic knowledge in general was performed by trained linguists through the careful, manual analysis of texts. This process is extremely time intensive and requires expertise. Therefore, more recent methods have leveraged machine-learning (ML) techniques to acquire linguistic knowledge, with the hope of reducing manual effort and dependence on linguistic expertise.

8.4.1 Morphology

Morphology concerns the combination of **morphemes** (roots, prefixes, suffixes) to produce words or lexemes, where a lexeme generally constitutes several forms of the same word (e.g. *activate*, *activates*, *activating*, *activated*, *activation*). **Free morphemes** can occur as separate words, while **bound morphemes** cannot (e.g. *de-* in *detoxify*, *-tion* in *creation*, *-s* in *dogs*). **Inflectional morphemes** express grammatically required features or indicate relations between different words in the sentence, but do not change the basic syntactic category; for example, *big*, *bigg-er*, *bigg-est* are all adjectives. **Derivational morphemes** change the part of speech or the basic meaning of a word: thus *-ation* added to a verb may form a noun (*activ-ation*) and *re-activate* means activate again. Biomedical language has a very rich morphological structure especially for chemicals (e.g. *Hydr-oxy-nitro-di-hydro-thym-ine*) and procedures (*hepatico-cholangio-jejuno-stom-y*), but recognizing morphemes is complex. In the previous chemical example, the first split must be made after *hydr-* (because the *-o-* is part of *–oxy*), while the fifth split occurs after *hydro-*. In the procedure example, an automated morphological analyzer would have to distinguish *stom* (mouth) from *tom* (cut) in *-stom*. NLP systems in the biomedical domain do not generally incorporate morphological knowledge. Instead many systems use **regular expressions** to represent what textual words consist of, and a **lexicon** to specify the words or lexemes in the domain and their linguistic characteristics.

Patterns are conveniently represented by the formalism known as a regular expression or equivalently, **finite state automata** (Jurafsky and Martin 2009, pp. 17–42). For example, the following simple regular expression represents what patterns are to be recognized as morphemes.

$$\mathbf{[a-z]+|[0-9]+}$$

The vertical bar (|) separates alternative expressions, which in this case specify two different kinds of **tokens** (alphabetic and numeric). Expressions in square brackets represent a range or choice of characters. The expression [a-z] indicates a lower case letter, while [0–9] indicates a digit. The plus sign denotes one or more occurrences of an expression. According to this regular expression, the sentence "*patient*'*s wbc dropped to 12*." contains six morphemes, and the string *patient*'*s* would contain two morphemes, *patient* and *s*; in this case, the *s* is a morpheme denoting a possessive construct. This regular expression is very limited, and would not represent a comprehensive set of tokens found in text. For example, it does not represent other morphological variations, such as negation (*n*'*t*) or numbers with a decimal point (*3.4*).

More complex regular expressions can handle many of the morphological phenomena described above. However, situations that are locally ambiguous are more challenging, and representations that can encode more knowledge are preferable. For instance, **Markov processes** (see Chap. 3) that encode some level of context by assigning probabilities to the presence of individual morphemes and to possible combinations of morphemes, can help characterize morphological knowledge. While an important field of study, there has been little work concerning morphology in the field of NLP in the biomedicine and health domains, especially for the English language. Encoding morphological knowledge is necessary in languages that are morphologically rich (e.g. Turkish, German, and Hebrew).

8.4.2 Syntax

Syntax concerns the categorization of the words in the language, and the structure of the phrases and sentences. Each word belongs to one or more **parts of speech** in the language, such as noun (e.g. *chest*), adjective (e.g. *mild*), or tensed verb (e.g. *improves*), which are the elementary components of the grammar and are generally specified in a **lexicon**. Words may also have subcategories, depending on the corresponding basic part of speech, which are usually expressed by inflectional morphemes. For example, nouns have number (e.g. plural or singular as in *legs*, *leg*), person (e.g. first, second, third as in *I*, *you*, *he*, respectively), and case (e.g. subjective, objective, possessive as in *I*, *me*, *my*, respectively). **Lexemes** can consist of more than one word as in foreign phrases (*ad hoc*), prepositions (*along with*), and idioms (*follow up*, *on and off*). Biomedical lexicons tend to contain many multiword lexemes, e.g. lexemes in the clinical domain include multiword terms such as *congestive heart* *failure* and *diabetes mellitus*, and in the biomolecular domain include the gene named *ALL1-fused gene from chromosome 1q*.

Lexemes combine (according to their parts of speech) in well-defined ways to form sequences of words or phrases, such as noun phrases (e.g. *severe chest pain*), adjectival phrases (e.g. *painful to touch*), or verb phrases (e.g. *has increased*). Each phrase generally consists of a main part of speech and modifiers, e.g. nouns are frequently modified by adjectives, while verbs are frequently modified by adverbs. The phrases then combine in well-defined ways to form sentences (e.g. "*he complained of severe chest pain*" *is a well-formed sentence*, *but* "*he severe complained chest pain of*" *is not*). General English imposes many restrictions on the formation of sentences, e.g. every sentence requires a verb, and count nouns (like *cough*) require an article (e.g., *a* or *the*). Clinical language, in contrast, is often **telegraphic**, relaxing many of these restrictions of the general language to achieve a highly compact form. For example, clinical language allows all of the following as sentences: "*the cough worsened*," "*cough worsening*," and "*cough*." Because the community widely uses and accepts these alternate forms, they are not considered ungrammatical, but constitute a **sublanguage** (Grishman and Kittredge 1986; Kittredge and Lehrberger 1982; Friedman 2002). There are a wide variety of sublanguages in the biomedical domain, each exhibiting specialized content and linguistic forms.

8.4.2.1 Representation of Syntactic Knowledge

Syntactic knowledge can be represented by means of a lexicon and a grammar. Representing such knowledge in a computable fashion is important, as it enables the creation of tools that parse the syntax of any given sentence, by matching the sentence against the lexicon and the grammar. A **lexicon** is used to delineate the lexemes along with their corresponding parts of speech and canonical forms so that each lexical entry assigns a word to one or more parts of speech, and also to a canonical form or lemma. For example, *abdominal* is an adjective where the canonical form is *abdomen*, and *activation* is a noun that is the nominal form of the verb *activate*. A **grammar** specifies the structure of the phrases and sentences in the language. It specifies how the words combine into well-formed structures through use of rules where categories combine with other categories or structures to produce a well-formed structure with underlying relations. A lexicon and grammar should be compatible with each other in that the parts of speech or categories specified in the lexicon should be the same as those specified in the rules of the grammar. In many systems, the parts of speech are typically standard and are based on the parts of speech specified by the Penn Treebank Project (Marcus et al. 1993). Table 8.1 provides some examples of Penn Treebank parts of speech, also called tags.

Generally, words combine to form phrases consisting of a **head word** and modifiers, and phrases combine to form sentences or clauses. For example, in English, there are noun phrases (NP) that contain a noun and optionally left and right modifiers, such as definite articles, adjectives, or prepositional phrases (i.e. *the patient*, *lower extremities*, *pain in lower extremities*, *chest pain*), and verb phrases (VP), such as *had pain*, *will be discharged*, and *denies smoking*. Phrases and sentences can be represented as a sequence where each word is followed by its corresponding part of speech. For example, S*evere joint pain* can be represented as *Severe/JJ joint/NN pain/NN*.

The field of computer science provides a number of formalisms that can be used to represent syntactic linguistic knowledge. These include symbolic or logical formalisms (e.g. **regular expressions** and **context free grammars**) and statistical formalisms (e.g. **probabilistic context free grammars**). Simple phrases, which are phrases that do not contain right modifiers, can be represented using **regular expressions**. When using regular expressions to represent the syntax of simple phrases, syntactic categories are used in the expression. An example of a regular expression (using the Penn Treebank parts of speech defined in Table 8.1) for a simple noun phrase is:

$$\mathbf{DT?JJ*NN*(NN\,|\,NNS)}$$

Table 8.1 Description of some part-of-speech tags from the Penn Treebank

Tag	Meaning (example)
CC	Conjunction (*and*)
CD	Cardinal number (*2*)
DT	Article (*the*)
IN	Preposition (*of*)
JJ	Adjective (*big*)
NN	Singular noun (*pain*)
NNS	Plural noun (*arms*)
PP$	Possessive pronoun (*her*)
PRP	Personal pronoun (*she*)
VB	Infinitive verb (*fall*)
VBD	Past-tense verb (*fell*)
VBG	Progressive verb form (*falling*)
VBN	Past participle (*fallen*)
VBZ	Present-tense verb (*falls*)

This structure specifies a simple noun phrase as consisting of an optional determiner (i.e. *a*, *the*, *some*, *no*), followed by zero or more adjectives, followed by zero or more singular nouns, and terminated by a singular or plural noun. For example, the above regular expression would match the noun phrase "*no/DT usual/JJ congestive/JJ heart/NN failure/NN symptoms/NNS*" but would not match "*heart/NN the/DT unusual/JJ*" because in the above regular expression *the* cannot occur in the middle of a noun phrase. In addition, the above regular expression does not cover many legitimate simple noun phrases, such as "*3/CD days/NNS*", "*her/PRP$ arm/NN*", and "*pain/NN and/CC fever/NN*".

A complex noun phrase with right modifiers, however, cannot be handled using a regular expression because a right modifier may contain **nested structures**, such as nested prepositional phrases or nested relative clauses. More complex language structures, like phrases with right modifiers, can be represented by a **context-free grammar** (CFG) or by equivalent formalisms (Jurafsky and Martin 2009, pp 386–421). A CFG is concerned with how sequences of words combine to form phrases or constituents. A very simple grammar of English is shown in Fig. 8.2. Context-free rules use **part-of-speech tags** (see Table 8.1) and the operators found in regular expressions. The difference is that each rule has a non-terminal symbol on the left side (S, NP, VP, PP) that represents a syntactic structure, and consists of a rule that specifies a sequence of grammar symbols (non-terminal and terminal) on the right side. Thus, in Fig. 8.2, the S (sentence) rule contains a sequence consisting of the symbols NP, followed by VP, which in turn are followed by a literal that is a '.'. Additionally, other rules may refer to these symbols or to the atomic parts of speech. In the NP rule there is an optional determiner DT as well as an optional prepositional phrase PP, which in turn contains an embedded NP to define noun phrases such as "*the pain*", "*pain*", "*pain in arm*", and "*pain in elbow of left arm*".

```
S       → NP VP .
NP      → DT? JJ* (NN | NNS) CONJN* PP* | NP and NP
VP      → (VBZ | VBP) NP? PP*
PP      → IN NP
CONJN → and (NN | NNS)
```

Fig. 8.2 A simple syntactic context-free grammar of English. A sentence is represented by the rule S, a noun phrase by the rule NP, a verb phrase by VP, and a prepositional phrase by PP. Terminal symbols in the grammar that correspond to syntactic parts of speech, are underlined in the figure

CFG grammar rules generally give rise to many possible structures for a parse tree, which represent sequences of alternative choices of rules in the grammar, but some choices are more likely than others. For example, in the sentence "*she experienced pain in chest*", it is more likely that *in chest* modifies *pain* and not *experienced*. The preferences can be represented using a **probabilistic context free grammar** that associates a probability with each choice in a rule (Jurafsky and Martin 2009, pp 459–479). The grammar shown in Fig. 8.2 can be augmented with probabilities for each rule (see Fig. 8.3). The number indicates the probability of including the given category in the parse tree. For example, the probability assigned to having a determiner (DT) in a NP can be 0.9 (and conversely not having a DT at the beginning of a NP has a probability of 0.1). The probability of a present tense verb (VBZ) is 0.4, while a past tense verb (VBD) is 0.6. The

$$\begin{array}{lll}
\mathrm{S} & \rightarrow & \mathrm{NP\ VP\ .} \\
\mathrm{NP} & \rightarrow & \mathrm{DT?}^{.9}\ \mathrm{JJ}^{*.8}\ (\mathrm{NN} \mid {}^{.6}\mathrm{NNS})\ \mathrm{PP}^{*.8} \\
\mathrm{VP} & \rightarrow & (\mathrm{VBZ} \mid {}^{.4}\mathrm{VBD})\ \mathrm{NP?}^{.9}\ \mathrm{PP}^{*.7} \\
\mathrm{PP} & \rightarrow & \underline{\mathrm{IN}}\ \mathrm{NP}
\end{array}$$

Fig. 8.3 A simple probabilistic context-free grammar of English. Probabilities for each rule are part of the grammar and are derived from a large corpus annotated with syntactic information

probabilities are usually estimated from a large corpus of text that has been annotated with the correct syntactic structures.

Recently, there has been increased interest in the **dependency grammar** formalism (Jurafsky and Martin 2009, 414–416), which focuses on how words relate to other words, although there are many different theories and forms of dependency grammars. Dependency is a binary asymmetrical relation between a head and its dependents or modifiers. The head of a sentence is usually a tensed verb. Thus, unlike CFGs, dependency structures do not contain phrasal structures but are basically directed relations between words. For example, in the sentence, *The patient had pain in lower extremities*, the head of the sentence is the verb *had*, which has two arguments, a subject noun *patient* and an object noun *extremities*, *the* modifies or is dependent on *patient*, and *in* is dependent on *pain*, *extremities* is dependent on *in*, and *lower* is dependent on *extremities*. As such, in a dependency grammar, the relations among words and the concept of head in particular (e.g. *extremities* is the head of *lower*) is closer to the semantics of a sentence. We introduce the concept of semantics next.

8.4.3 Semantics

Semantics concerns the meaning or interpretation of words, phrases and sentences, generally associated with real-world applications. There are many different theories for representation of meaning, such as **logic-based**, **frame-based**, or **conceptual graph** formalisms. In this section, we discuss semantics involved in interpretation of the text in order to accomplish practical tasks, and do not aim to discuss representation of complete meaning. Each word has one or more meanings or **word senses** (e.g. *capsule*, as in *renal capsule*, *vitamin B12 capsule*, or *shoulder capsule*), and other terms may modify the senses (e.g. *no*, as in *no fever*, or *last week* as in *fever last week*). Additionally, the meanings of the words combine to form a meaningful sentence, as in *there was thickening in the renal capsule*). Representation of the semantics of general language is extremely important, but the underlying concepts are not as clear or uniform as those concerning syntax. Interpreting the meaning of words and text for general language is very challenging, but interpreting the meaning of text within a sublanguage is more feasible. More specifically, biomedical sublanguages are easier to interpret than general languages because they exhibit more restrictive semantic patterns that can be represented more easily (Harris et al. 1989; Harris 1991; Sager et al. 1987). Sublanguages tend to have a relatively small number of well-defined **semantic types** (e.g. medication, gene, disease, body part, or organism) and a small number of **semantic patterns** (e.g. medication-treats-disease, gene-interacts with-gene).

8.4.3.1 Representation of Semantic Knowledge

Semantic interpretations must be assigned to individual terms (e.g. single words or multi-word terms that function as single words) that are then combined into larger semantic structures (Jurafsky and Martin 2009, pp 545–580). The notion of what constitutes an individual term is not straightforward, particularly in the biomedical domain where there are many multi-word terms that have specific meanings that are different from the combined meanings of the parts. For example, *burn* in *heart burn* has a different meaning than in *finger burn*. Semantic information about words can be maintained in a lexicon or a domain-specific dictionary. A **semantic type or class** is usually a broad class that is associated with a specific domain and includes many instances, while a **semantic sense** distinguishes individual word meanings (Jurafsky and Martin 2009, pp 611–617). For example, *heart attack*, *myocardial infarct*, and *systemic lupus* *eryhtematosus* (*SLE*) all have the same semantic type (disease); *heart attack* and *myocardial infarct* share the same semantic sense (they are synonymous) that is distinct from the sense of *SLE* (a different condition).

A lexicon containing semantic knowledge may be created manually by a linguist, or be derived from external knowledge sources. Examples of semantic knowledge sources are the **Unified Medical Language System** (UMLS) (Lindberg et al. 1993), which assigns semantic types to terms, such as **disease**, **procedure**, or **medication**, and also specifies the sense of the concept through a unique concept identifier (CUI), and GenBank (Benson et al. 2003), which lists the names of genes and also specifies a unique identifier for each gene concept. While external sources can save a substantial effort, the types and senses provided may not be the appropriate granularity for the text being analyzed. Narrow categories may be too restrictive, and broad categories may introduce ambiguities. Morphological knowledge can be helpful in determining semantic types in the absence of lexical information. For example, in the clinical domain, suffixes like *–itis* and *-osis* indicate diseases, while *-otomy* and *ectomy* indicate procedures. However, such techniques cannot determine the specific sense of a word.

Semantic structures consisting of **semantic relations** can be identified using regular expressions, which specify patterns of semantic types that are relevant in a specific domain, and that are associated with a real-word interpretation. The expressions may be semantic and look only at the semantic categories of the words in the sentence. For example, this method may be applied in the biomolecular domain to identify interactions between genes or proteins. For example, the regular expression

$$[\mathbf{GENE}\,|\,\mathbf{PROT}]\mathbf{MFUN}[\mathbf{GENE}\,|\,\mathbf{PROT}]$$

will match sentences consisting of very simple gene or protein interactions (e.g. *Pax-3/GENE activated/MFUN Myod/GENE*). In this case, the elements of the pattern consist of semantic classes: gene (GENE), molecular function (MFUN), and protein (PROT). This pattern is very restrictive, and regular expressions that skip over parts of the sentence can be written to detect relevant patterns for a broader variety of text, although it will incur some loss of specificity and precision while achieving increased sensitivity. For example, the regular expression

$$[\mathbf{GENE}\,|\,\mathbf{PROT}].{*}\,\mathbf{MFUN}.{*}[\mathbf{GENE}\,|\,\mathbf{PROT}]$$

specifies a pattern that provides for skipping over terms in the text. The dot (.) matches any tag, and the asterisk (*) allows for an arbitrary number of occurrences. Using the above expression, the interpretation of the interaction, *Pax-3 activated Myod* would be obtained for the sentence "*Pax-3/GENE, only when activated/MFUN by Myod/GENE, inhibited/MFUN phosphorylation/MFUN*". In this example, the match does not capture the information correctly because "*only when*" was skipped. The correct interpretation of the individual interactions in this sentence should be "*Myod activated Pax-3*", and "*Pax-3 inhibited phosphorylation*". Note that the simple regular expression shown above does not provide for the latter pattern (i.e. GENE-MFUN-MFUN), for the connective relation *only when*, or for the passive structure *activated by*.

Frequently, simple semantic relations are not represented using regular expressions. Instead, relations and the roles of the elements in the relation are specified by manual or semi-automated annotation so that machine-learning methods could be leveraged to develop models that detect the relations. For example, interaction type relations could be annotated so that it has an element that is an agent, and an element that is a target. Therefore, the above sentence would be annotated as consisting of two relations: an interaction relation *inhibit* where the agent is *Pax-3* and the target is *phosphorylation*, and an interaction relation *activate* where the agent is *Myod* and the target is *Pax-3*.

More complex semantic structures containing nesting can be represented using a **semantic grammar**, which is a context free grammar based on semantic categories. As shown in Fig. 8.4, a simple semantic grammar for clinical texts might define a clinical sentence as a Finding, which

S	→	Finding .
Finding	→	DegreePhrase? ChangePhrase? SYMP
ChangePhrase	→	NEG? CHNG
DegreePhrase	→	DEGR ∣ NEG

Fig. 8.4 A simple semantic context-free grammar for the English clinical domain. A sentence S consists of a Finding that consists of an optional DegreePhrase, an optional ChangePhrase and a Symptom. The DegreePhrase consists of a degree type word or a negation type word; the ChangePhrase consists of an optional negation type word followed by a change type word. The terminal symbols in the grammar correspond to semantic parts of speech and are underlined

consists of optional degree information and optional change information followed by a symptom. Such a semantic grammar can parse the sentence "increased/CHNG tenderness/SYMP", a typical sentence in the clinical domain, which often omits subjects and verbs.

NLP systems can represent more complex language structures by integrating syntactic and semantic structures into the grammar (Friedman et al. 1994). In this case, the grammar would be similar to that shown in Fig. 8.4, but the rules would also include syntactic structures. Additionally, the grammar rule may also specify the representational output form that represents the underlying interpretation of the relations. For example, in Fig. 8.4, the rule for Finding would specify an output form denoting that SYMP is the primary finding and the other elements are the modifiers.

More comprehensive syntactic structures can be recognized using a broad-coverage CFG of English that is subsequently combined with a semantic component (Sager et al. 1987). After the syntactic structures are recognized, they are followed by syntactic rules that regularize the structures. For example, passive sentences, such as "*the chest x-ray was interpreted by a radiologist*", would be transformed to the active form (e.g. "*a radiologist interpreted the chest x-ray*"). Another set of semantic rules would then operate on the regularized syntactic structures to interpret their semantic relations.

The representation of a probabilistic CFG that contains semantic information would be similar to that of a CFG for syntax, but it would require a large corpus that has been annotated with both syntactic and semantic information. Since a semantic grammar is domain and/or application specific, annotation involving the phrase structure would be costly and not portable, and therefore is not generally done.

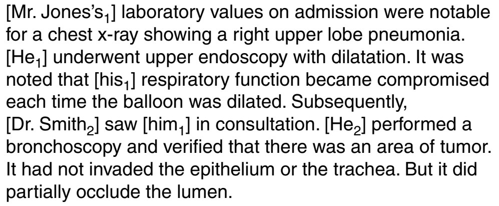
8.4.4 Pragmatics and Discourse

Pragmatics concerns how knowledge concerning the intent of the author of the text, or more generally the context in which the text is written, influences the meaning of a sentence or a text. For example, in a mammography report *mass* generally denotes *breast mass*, whereas a radiological report of the chest denotes *mass in lung*. In yet a different genre of texts, like a religious journal, it is likely to denote a ceremony. Similarly, in a health care setting, *he drinks heavily*, is assumed to be referring to alcohol and not water. In these two examples, pragmatics influences the meaning of individual words. It can also influence the meaning of larger linguistic units. For instance, when physicians document the chief-complaint section of a note, they list symptoms and signs, as reported by the patient. The presence of a particular symptom, however, does not imply that the patient actually has the symptom. Rather, it is understood implicitly by both the author of the note and its reader that this is the patient's impression rather than the truth. Thus, the meaning of the chief-complaint section of a note is quite different from the assessment and plan, for instance. Another pragmatic consideration is the interpretation of pronouns and other referential expressions (*there*, *tomorrow*). For example, in the two following sentences "*An infiltrate was noted in right upper lobe. It was patchy*", the pronoun *it* refers to *infiltrate* and not *lobe*. In a sentence containing the term *tomorrow*, it would be necessary to know when the note was written in order to interpret the actual date denoted by *tomorrow*.

In the biomedical domain, pragmatics is taken into account, but knowledge about the pragmatics of the domain is not modeled explicitly. In NLP applications limited to a specific subdomain, it can be encoded through the semantic lexicon (like the one described in Sect. 8.4.3) and rules about the discourse of a text.

While sentences in isolation convey individual pieces of information, sentences together combine to form a text, obeying a **discourse** structure (e.g. a group of sentences about the same topic can be grouped into coherent paragraphs, whereas a dialogue between a physician and a patient can be structured as a sequence of conversation turns). Complete analysis of a text requires analysis of relationships between sentences and larger units of discourse, such as paragraphs and sections (Jurafsky and Martin 2009, pp 681–723).

There has been much work regarding discourse in computational linguistics and in NLP in the general domain. One of the most important mechanisms in language for creating linkages between sentences is the use of **referential expressions**, which include pronouns (*he*, *she*, *her*, *himself*), proper nouns (*Dr. Smith*, *Atlantic Hospital*) and noun phrases modified by the definite article or a demonstrative (*the left breast*, *this medication*, *that day*, *these findings*). **Coreference chains** provide a compact representation for encoding the words and phrases in a text that all refer to the same entity. Figure 8.5 shows a text and the coreference chains corresponding to two entities in the discourse. Each referential expression has a unique referent that must be identified in order to make sense of the text. In the figure, the proper noun *Dr. Smith* refers to the physician who is treating the patient. In the first two sentences, *his* and *he* refer to the patient, while *he* refers to the physician in the fourth sentence. In that figure, there also are several definite noun phrases (e.g. *the epithelium*, *the trachea*, and *the lumen*) that have to be resolved. In this case, the referents are parts of the patient's body and are not mentioned previously in the text.



[Mr. Jones's$_1$] laboratory values on admission were notable for a chest x-ray showing a right upper lobe pneumonia. [He$_1$] underwent upper endoscopy with dilatation. It was noted that [his$_1$] respiratory function became compromised each time the balloon was dilated. Subsequently, [Dr. Smith$_2$] saw [him$_1$] in consultation. [He$_2$] performed a bronchoscopy and verified that there was an area of tumor. It had not invaded the epithelium or the trachea. But it did partially occlude the lumen.

Fig. 8.5 A text and two coreference chains, one about the patient and one about the clinician. *Brackets* denotes the span of the reference, and the indices inside the *brackets* refer to the entity. Note that even though we show two chains only, there are many other entities in the text (such as "laboratory values," "chest x-ray, and "tumor")

8.5 NLP Techniques

Most NLP systems are designed with separate modules that handle different functions. The modules typically roughly coincide with the linguistic levels described in Sect. 8.4. In general, the output from each lower level serves as input to the next higher level. For example, the output of **tokenization** transforms a textual string into tokens that will undergo lexical analysis to determine their parts of speech and possibly other properties as specified in a lexicon; the parts of speech along with the corresponding lexical definitions will then be the input to syntactic analysis that will determine the structure of the sentence; the structure will be the input to semantic analysis that will interpret the meaning. Each system packages these processing steps somewhat differently. At each stage of processing, the module for that stage aims to regularize the data in some aspect to reduce variety while preserving the informational content as much as possible.

8.5.1 Low-Level Text Processing

In addition to different linguistic levels introduced in the previous section, there is an additional practical level involved in NLP that pertains to the low-level processing of an input file. **Text processing**, as it is commonly referred to, is a sometimes-tedious part of developing an NLP system, but it is a critical one, as it impacts the processing at all the subsequent linguistic levels. Below, we review a few of the characteristics of low-level text that need to be considered when implementing an NLP system or when preparing input for an off-the-shelf NLP system.

8.5.1.1 File Formats

Documents in a corpus (a collection of documents, plural: corpora) can be stored using different formats. In some EHRs, for instance, it is not uncommon to have patient notes stored in **Rich Text Format** (RTF). PubMed citations can be downloaded from the Web in an **Extensible Markup Language** (XML) format, whereas full-text articles from PubMed can be **Hypertext Markup Language** (HTML) files, and Twitter feeds can come in the XML-based **Really Simple Syndication** (RSS) format. While there is no established way to convert one format into another or to clean up HTML tags, there are several packages available in many programming languages that can help deal with different file formats.

8.5.1.2 Character Sets and Encodings

Characters can be encoded in different ways in a computer (e.g. ASCII, which contains 128 characters, or Unicode, which contains over 100,000 characters). Knowing the character encoding used in a document is essential to recognize the characters and process the text further.

8.5.2 Document Structure

Journal articles generally have well-defined section headers, associated with the informational content (e.g. Introduction, Background, Methods, Results) and other information, such as references and data about the authors and their affiliations. In clinical reports, there often are well-defined sections (e.g. History of Present Illness, Past Medical History, Family History, Allergies, Medications, Assessment and Plan) that would be important for an NLP system to recognize. For example, medications mentioned in the Medication Section have been prescribed to patients whereas medications mentioned in the Allergy Section should not be. Within sections, other document-level constructs must be identified, such as paragraphs, tables in various formats, and (often nested) list items. In many cases however, a text comes "raw," that is, without any formatting information, or with some idiosyncratic formatting. In the clinical domain, for instance, there is evidence that many typed notes have no explicit section headers.

Tokens. The next step in processing generally consists of separating the cleaned up ASCII text, which is usually one large string at this stage, into individual units called tokens (a process called tokenization), which include morphemes, words (often morpheme sequences), numbers, symbols (e.g. mathematical operators), and punctuation. The notion of what constitutes a token is far from trivial. The primary indication of a token in general English is the occurrence of white space before and after it; however there are many exceptions: a token may be followed by certain punctuation marks without an intervening space, such as by a period, comma, semicolon, or question mark, or may have a '-'in the middle. In biomedicine, periods and other punctuation marks can be part of words (e.g. *q.d.* meaning *every day* in the clinical domain or *M03F4.2A*, a gene name that includes a period), and are used inconsistently, thereby complicating the tokenization process. For instance, in the string "*5 mg. given*" the tokenization process determines whether to keep the string "*mg*." as one token or two ("*mg*" and "."). In addition, chemical and biological names often include parentheses, commas and hyphens (for example, "(*w*)*adh-2*") that also complicate the tokenization process.

Symbolic approaches to tokenization are generally based on pattern matching using regular expressions, but most current approaches use statistical methods. One method consists of comparing which alternatives are most frequent in a correctly-tokenized corpus (e.g. is the token *M03F4.2A* more frequent than the two tokens *M03F4* and *2A* one next to another, but separated by a white space).

8.5.2.1 Sentence Boundaries

Detecting the beginning and end of a sentence may seem like an easy task, but is a highly domain-dependent one. Not all sentences end with a punctuation mark (this is especially true in texts with minimal editing, such as online patient posts and clinical notes entered by physicians). Conversely, the presence of a period does not necessarily mean the end of a sentence (as discussed in tokenization). While there are statistically trained models available for general NLP, they are based on an annotated corpus of general language text, and do not perform as well on biomedical and health-related text. Therefore, many NLP systems rely on hand-built rules to detect an end of sentence.

8.5.2.2 Case

Most NLP techniques operate on words as their smallest unit of processing. The definition of a word is not trivial, however. One question is whether to consider strings with different cases the same. In some situations, it makes sense to keep all tokens in lowercase, as it reduces variations in vocabulary. But in others, it might hinder further NLP: there are many acronyms that, when lowercased, might be confused with regular words, such as *TEN*, which is an abbreviation for *toxic epidermal necrosis*.

8.5.3 Syntax

There are generally two tasks involved in syntactic analysis: one involves determining the syntactic categories or parts of speech of the words, and the other involves determining and representing the structures of the sentences.

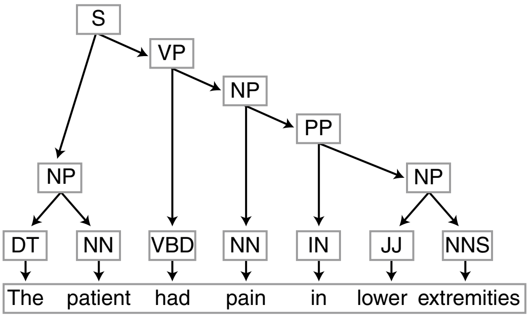


Fig. 8.6 A parse tree for the sentence *the patient had pain in lower extremities* according to the context-free grammar shown in Fig. 8.2. Notice that the terminal nodes in the tree correspond to the syntactic categories of the words in the sentence

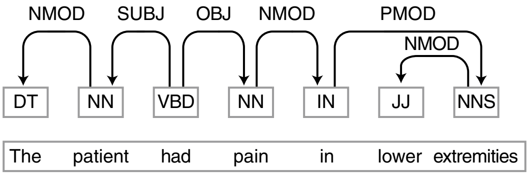


Fig. 8.7 A parse tree for the sentence *the patient had pain in lower extremities* in a dependency grammar framework

8.5.3.1 Output of Syntactic Parse

Applying grammar rules to a given sentence is called parsing, and if the grammar rules can be satisfied, the grammar yields a nested structure that can be represented graphically as a **parse tree**. For example, based on the CFG shown in Fig. 8.2, the sentence "*the patient had pain in lower extremities*" would be parsed successfully and would be assigned the parse tree shown in Fig. 8.6. Alternatively, brackets can be used to represent the nesting of phrases instead of a parse tree. Subscripts on the brackets specify the type of phrase or tag:

$$[_{S}\,[_{NP}\,[_{DT}\,\textbf{the}]\,[_{NN}\,\textbf{patient}]]\,[_{VP}\,[_{VBD}\,\textbf{had}]$$
$$[_{NP}\,[_{NN}\,\textbf{pain}][_{PP}[_{IN}\,\textbf{in}]\,\,[_{NP}\,[_{JJ}\,\textbf{lower}][_{NNS}\,\textbf{extremities}]]]]]]$$

The following shows an example of a parse in the biomolecular domain for the sentence *Activation of Pax-3 blocks Myod phosphorylation*:

$$[_{S}[_{NP}[_{NN}\,\textbf{Activation}]\,[_{PP}[_{IN}\,\textbf{of}\,]\,[_{NP}\,[_{NN}\,\textbf{Pax-3}]]]]$$
$$[_{VP}[_{VBZ}\,\textbf{blocks}]\,[_{NP}[_{JJ}\,\textbf{Myod}]\,[_{NN}\,\textbf{phosphorylation}]]]]$$

When a dependency grammar is used, the representation of the parsed structure would be different from the structure generated by a CFG, and would reflect relations between words instead of structures. Figure 8.7 shows a representation of a dependency structure for the same sentence as the one illustrated in Fig. 8.6. In this example, the noun *patient* and the noun *pain* are the subject and object arguments of the verb *had*, and thus are both dependent on the verb. Similarly, the determiner *the* modifies (i.e. is dependent on) *patient*, and the preposition *in* modifies *pain*.

8.5.3.2 Part-of-Speech Tagging and Lexical Lookup

Once text is tokenized, an NLP system needs to identify the words or multi-word terms known to the system, and determine their categories and canonical forms. Many systems carry out tokenization on complete words and perform **part-of-speech tagging**, and then some form of lexical look up immediately afterwards. This requires that the tagger and lexicon contain all the possible combinations of morphemes. A few systems perform morphological analysis during tokenization. In that case, the lexicon only needs entries for roots, prefixes, and suffixes, with additional entries for irregular forms. For example, the lexicon would contain entries for the roots *abdomen* (with variant *abdomin-*) and *activat-*, the adjective suffix *-al*, verb suffix *-e*, and noun suffix *-ion*.

Part-of-speech tagging is not straightforward because a word may be associated with more than one part of speech. For example, *stay* may be a noun (as in *her hospital stay*) or a verb (as in *refused to stay*). Without resolution, such ambiguities can lead to creating different structures for the sentence causing inaccuracies in parsing and interpretation, resulting in a substantial decrease in performance. For example, when *eating* occurs before a noun, it can be an adjective (JJ) that modifies the noun, or it can be a verb form (VBG) with the noun as object:

She/PRP denied/VBD eating/JJ difficulties/NN
She/PRP denied/VBD eating/VBG food/NN

In the first example, the patient is having a difficulty associated with eating, whereas in the second the patient is denying the act of eating food. Various methods for part-of-speech tagging may be used to resolve ambiguities by considering the surrounding words. A rule-based approach generally consists of rules based on the word that precedes or follows the current word. For example, if *stay* follows *the* or *her*, a rule may specify that it should be tagged as a noun, but if it follows *to* it should be tagged as a verb.

Currently, the most widely used approaches are statistically based part-of-speech taggers. One type of approach is based on Markov models (as described above for morphology). In this case, the **transition matrix** specifies the probability of one part of speech following another (see Table 8.2):

Table 8.2 Transition probabilities for part-of-speech tags

	NN	VB	VBD	VBN	TO	IN
NN	0.34	0.00	0.22	0.02	0.01	0.40
VB	0.28	0.01	0.02	0.27	0.04	0.39
VBD	0.12	0.01	0.01	0.62	0.05	0.19
VBN	0.21	0.00	0.00	0.03	0.11	0.65
TO	0.02	0.98	0.00	0.00	0.00	0.00
IN	0.85	0.00	0.02	0.05	0.00	0.08

The following sentence shows the correct assignment of part-of-speech tags: *Rheumatology/NN consult/NN continued/VBD to/TO follow/VB patient/NN.*

Table 8.3 Probabilities of alternative part-of-speech tag sequences

Part of speech tag sequence	Probability
NN NN VBD TO VB NN	0.001149434
NN NN VBN TO VB NN	0.000187779
NN VB VBN TO VB NN	0.000014194
NN NN VBD IN VB NN	0.000005510
NN NN VBN IN VB NN	0.000001619
NN VB VBD TO VB NN	0.000000453
NN VB VBN IN VB NN	0.000000122
NN VB VBD IN VB NN	0.000000002

This assignment is challenging for a computer, because *consult* can be tagged VB (*Orthopedics asked to consult*), *continued* can be tagged VBN (*penicillin was continued*), and *to* can be tagged IN. However, probabilities can be calculated for these sequences using the matrix in Table 8.2 (these were estimated from a large corpus of clinical text). By multiplying the transitions together, a probability for each sequence can be obtained (as described above for morphology), and is shown in Table 8.3. Note that the correct assignment has the highest probability.

8.5.3.3 Parsing

Many NLP systems perform some type of syntactic parsing to determine the structure of the sentence. Some systems perform **partial parsing** or **shallow parsing**, and are based on a number of different methods (Jurafsky and Martin 2009; pp 450–458). Partial parsing systems determine the structure of local phrases, such as simple noun phrases (i.e. noun phrases without right adjuncts) and simple adjectival phrases but do not determine relations among the phrases. Determining non-recursive phrases where each phrase corresponds to a specific part of speech is also called

chunking. These systems tend to be robust because it is easier to recognize isolated phrases than it is to recognize complete sentences, but typically they lose some information. For example, in *amputation below knee*, the two noun phrases *amputation* and *knee* would be extracted, but the relation *below* might not be. Rule-based methods use regular expressions that are manually created, while statistical models are developed based on an annotated corpus.

Complete syntactic parsing recognizes and determines the structure of a complete sentence. A system based on a CFG tries to find a match between the sentence and the grammar rules. Therefore, parsing can be thought of as a search problem that tries to fit the sequence of part-of-speech tags associated with the sentence to all possible combinations of grammar rules. There are a number of different approaches to parsing, such as searching through the rules in a **top-down** or **bottom-up** fashion, or using dynamic programming methods for efficiency, as in the **Cocke–Younger–Kasami** (CYK), **Earley**, or **chart parsing** methods (Jurafsky and Martin 2009; pp 427–450).

Additionally, grammar rules generally give rise to many possible structures for a parse tree (structural ambiguity) because the sequence of alternative choices of rules in the grammar can yield different groupings of phrases based on syntax alone. For example, sentence 1a below corresponds to a parse based on the grammar rules shown in Fig. 8.2 where the VP rule contains a PP (e.g. *denied in the ER*) and the NP rule contains only a noun (e.g. *pain*). Sentence 1b corresponds to the same atomic sequence of syntactic categories but the parse is different because the VP rule contains only a verb (e.g. *denied*) and the NP contains a noun followed by a PP (e.g. *pain in the abdomen*). Prepositions and conjunctions are also a frequent cause of ambiguity. In 2a, the NP consists of a conjunction of the head nouns so that the left adjunct (e.g. *pulmonary*) is distributed across both nouns (i.e. this is equivalent to an interpretation *pulmonary edema* and *pulmonary effusion*), whereas in 2b, the left adjunct *pulmonary* is attached only to *edema* and is not related to *anemia*. In 3a the NP in the prepositional phrase PP contains a conjunction (i.e. this is equivalent to *pain in hands* and *pain in feet*) whereas in 3b two NPs are also conjoined but the first NP consists of *pain in hands* and the second consists of *fever*.

1a. *Denied* [*pain*] [*in the ER*]
1b. *Denied* [*pain* [*in the abdomen*]]
2a. *Pulmonary* [*edema and effusion*]
2b. [*Pulmonary edema*] *and anemia*
3a. *Pain in* [*hands and feet*]
3b. [*Pain in hands*] *and fever*

More complex forms of ambiguity do not exhibit differences in parts of speech or in grouping, but require determining deeper syntactic relationships. For example, when a verb ending in *–ing* is followed by *of*, the following noun can be either the subject or object of the verb.

Feeling of lightheadedness improved.
Feeling of patient improved.

Statistical approaches provide one method of addressing parsing ambiguity, and provide a mechanism where it is possible to prefer the more likely parses over less likely ones based on the probability of a parse tree that is the product of the probabilities of each grammar rule used to make it. For example, there are two ways to parse *x-ray shows patches in lung* using this grammar (shown below). The first interpretation in which *shows* is modified by *lung* has probability 3.48 × 10^{-8}, while the second interpretation in which *patches* is modified by *lung* has probability 5.97 × 10^{-8}.

$$\mathbf{[_{S}\ [_{NP}\ NN\ 0.1\times0.2\times0.6\times0.2]\,[_{VP}\ VBZ\ [_{NP}\ NN\ 0.1\times0.2\times0.6\times0.2]}$$
$$\mathbf{[_{PP}\ IN\ [_{NP}\ NN\ 0.1\times0.2\times0.6\times0.2]]\ \ 0.4\times.9\times.7\,]]}$$

$$\mathbf{[_{S}\ [_{NP}\ NN\ 0.1\times0.2\times0.6\times0.2\,]}$$
$$\mathbf{[_{VP}\ VBZ\ \ [_{NP}\ [_{PP}\ IN\ [_{NP}\ NN\ 0.1\times0.2\times0.6\times0.2\,]\ NN\ 0.1\times0.2\times0.6\times0.8\,]\ 0.4\times0.9\times0.3\,]]}$$

The probabilities are generally established based on a large corpus that has been parsed and annotated correctly. Therefore, it is critical that the corpus be of a genre similar to the text to be parsed; otherwise, performance will likely deteriorate.

8.5.4 Semantics

Semantic analysis involves steps analogous to those described above for syntax. First, semantic interpretations must be assigned to individual words, and then, these are combined into larger semantic structures, including modifications and relations. There are a number of formalisms for representing information in language (see Jurafsky and Martin 2009; pp 545–580), such as database tables, XML, **frames**, **conceptual graphs**, or **predicate logic**. Below, we discuss a few general approaches within this domain.

NLP systems may capture the clinical information at many different levels of granularity. One level of coarse granularity consists of classification of reports. For example, several systems (Aronow et al. 1995; Aronow et al. 1999) classified reports as positive or negative for specific clinical conditions, such as breast cancer. Another level of granularity that is useful for information retrieval and indexing, captures relevant terms by mapping the information to a controlled vocabulary, such as the UMLS (Aronson 2001; Nadkarni et al. 2001), but modifier relations are not captured. A more specific level of granularity also captures positive and negative modification (Mutalik et al. 2001; Chapman et al. 2001) or temporal status (Harkema et al. 2009). An even more specific level of granularity captures a comprehensive set of modifiers associated with the term, facilitating reliable information extraction (Friedman et al. 2004).

8.5.4.1 Output of Semantic Interpretation

Typically within a specific domain, semantic interpretation is limited in the sense that NLP systems in that domain do not attempt to capture the complete meaning of information in the text, but instead aim to capture limited relevant elements of the text. This process involves several representational aspects: word sense, relevant information that modifies or changes the underlying meaning of the word senses, well-defined relations in the domain, and relevant information that modifies the relations.

Representation of word senses is usually achieved by means of the many controlled vocabularies or **ontologies** in the domain that associate words or terms with unique meanings, or concepts, and have corresponding codes. Therefore, semantic interpretation of a word or term entails mapping it to a code that represents its concept. As with parts of speech, many words have more than one semantic interpretation or sense; an NLP system must determine which of these is intended in the given context and map the sense to a well-defined concept. For example, *growth* can be either an abnormal physiologic process (e.g. for a tumor) or a normal one (e.g. for a child). The word *left* can indicate laterality (*pain in left leg*) or an action (*patient left hospital*).

Relations between words may be represented a number of different ways. One very general representational form is a frame-based representation in which the **frame** contains **slots** consisting of predefined types of information, which may be optional (Minsky 1975). In addition, relations between the slots are predetermined. Thus, a frame may be used to represent a simple concept along with its modifiers. For example, a frame representing a patient's condition, could have a slot for the condition, such as *cough*, and additional slots that modify *cough*, such as a slot for severity as in *severe cough*, a slot for temporal information, as in *cough for 4 days*, a slot for type of cough, as in *productive cough*, and a slot for negation, as in *denies cough* (Friedman et al. 2004). A frame may also contain more complex information associated with relations between entities. For example, an interaction frame could be defined to contain a slot for the interaction, a slot for the agent, a slot for the target, and slots that modify the relation, such as a negation slot to represent an interaction that is negated, or a degree slot to represent the strength of an interaction.

Another representation could be a predicate-argument form where the interpretation of the **predicate** and the roles of the **arguments** are specified. Representation of an interaction between biomolecular substances could be *Interaction*(*Agent*,*Target*). For example, *Wnt blocks Pax-3* could be represented as *block*(*Wnt*,*Pax-3*) or the arguments could be codes uniquely defining those substances. Similarly, relations between medications and conditions can be represented as a predicate, such as *treat*, *prevent*, or *causes*, where the first argument could be the medication or a corresponding code, and the second argument could be the condition or its code.

8.5.4.2 Word Sense Interpretation

The problem of determining the correct sense of a term can involve matching a word to a concept in an ontology or controlled vocabulary in the given domain. For example, the UMLS identifies unique concepts, along with their variant and synonymous forms, and it would be straightforward to match the word *cough* in text to the concept cough (i.e. C0010200) in the UMLS. However, a word may have more than one sense, and the process of determining the correct sense is called **word sense disambiguation** (WSD). For example, *MS* could be mapped to C0026269, which corresponds to *mitral stenosis*, or to C0026769, which corresponds to *multiple sclerosis*. However, the underlying interpretation may not be in the UMLS, as in the sense *Ms* (the honorific title). WSD is much harder than syntactic disambiguation because there is no well-established notion of word sense, different lexicons or ontologies recognize different distinctions, and the space of word senses is substantially larger than that of syntactic categories. Words may be ambiguous within a particular domain, across domains, or in the natural language (e.g., English) in general. Abbreviations are notoriously ambiguous. The ambiguity problem is particularly troublesome in the biomolecular domain because biomolecular symbols in many model organism databases consist of three letters, and are ambiguous with other English words, and also with different gene symbols of different model organisms. For example, *nervous* and *to* are English words that are also the names of genes. When writing about a specific organism, authors use alias names that may correspond to different genes. For example, in articles associated with the mouse, according to the Mouse Genome Database (MGD) (Blake et al. 2003), authors may use the term *fbp1* to denote different genes.

Semantic disambiguation of lexemes can be performed using similar rule-based or statistical methods described above for syntax. Rules can assign semantic types using contextual knowledge of other nearby words and their types. For example, *discharge from hospital* and *discharge from eye* can be disambiguated depending on whether the noun following *discharge* is an institution or a body location. However, statistical approaches are generally used to determine the most likely assignment of semantic types based on contextual information, such as surrounding words (Jurafsky and Martin 2009, pp 637–667). As with statistical methods for morphology and syntax, large amounts of training data are required to provide sufficient instances of the different senses for each ambiguous word. This is extremely labor intensive because it means that a large manually annotated corpus for each ambiguous word must be collected where optimally each sense occurs numerous times, although in certain cases automated annotation is possible.

8.5.4.3 Interpretation of Relations among Words

Similar to syntactic analysis of simple phrases, semantic analysis can also be achieved using regular expressions. An alternate, robust method for processing sentences with regular expressions, which is often employed in general, uses **cascading finite state automatas** (**FSAs**) (Hobbs et al. 1996). In this technique, a series of different FSAs are employed so that each performs a special tagging function. The tagged output of one FSA becomes the input to a subsequent FSA. For example, one FSA may perform tokenization and lexical lookup, another may perform partial parsing to identify syntactic phrases, such as noun phrases and verb phrases, the next may perform named entity recognition (NER), and the next may recognize semantic relations, determine the roles of the entities, and map the entities to a predicate-argument or frame-based structure. In some cases, the patterns for the semantic relations will be based on a combination of syntactic phrases and their corresponding semantic classes, as shown below. The pattern for biomolecular interactions might then be represented using a combination of tags:

$$\mathbf{NP}_{[\mathrm{GENE|PROT}]}.^{*}\,\mathbf{VP}_{\mathrm{MFUN}}.^{*}\,\mathbf{NP}_{[\mathrm{GENE|PROT}]}$$

The advantage of cascading FSA systems is that they are relatively easy to adapt to different information extraction tasks because the FSAs that are domain independent (tokenizing and phrasal FSAs) remain the same while the domain-specific components (Semantic patterns) change with the domain and or the extraction task. These types of systems have been widely used in the 1990s to extract highly specific information, such as detection of terrorist attacks, identification of joint mergers, and changes in corporation management (Sundheim 1991, 1992, 1994, 1996, Grishman and Sundheim 1996). They are generally used to extract information from the literature, but they may not be accurate enough for clinical applications.

Complete parsing, similar to the methods used for syntactic parsing, can be used in systems that have a context-free semantic grammar. For example, the sentence *No increased tenderness* would be parsed correctly using the simple grammar shown in Fig. 8.4, where there is a finding that consists of a changephrase that has a negation (e.g. *no*) that modifies the change (e.g. *increased*), and the changephrase is followed by a symptom (e.g. *tenderness*). Therefore, *no* modifies the change and not the symptom. Note that ambiguity is possible in this grammar because a sentence such as *No/NEG increased/CHNG tenderness/SYMP* could be parsed by satisfying other rules. In an incorrect parse, the degreephrase (e.g. *no*) and the changephrase (e.g. *increased*) both modify *tenderness*; in this interpretation the symptom is negated and the change information is not.

8.5.5 Discourse

We focus in this section on one particular task as an example of discourse processing: automated resolution of referential expressions.

8.5.5.1 Automated Resolution of Referential Expressions

Coreference resolution can draw on both syntactic and semantic information in the text. Syntactic information for resolving referential expressions includes:

- Agreement of syntactic features between the referential phrase and potential referents
- Recency of potential referents (nearness to referential phrase)
- Syntactic position of potential referents (e.g. subject, direct object, object of preposition)
- The pattern of transitions of topics across the sentences

Syntactic features that aid the resolution include such distinctions as singular/plural, animate/inanimate, and subjective/objective/possessive. For example, pronouns in the text in Fig. 8.5 carry the following features: *he* (singular, animate, subjective), *his* (singular, animate, possessive) and *it* (singular, inanimate, subjective/objective). Animate pronouns (*he*, *she*, *her*) almost always refer humans. The inanimate pronoun *it* usually refers to things (e.g. *it had not invaded*), but sometimes does not refer to anything when it occurs in "cleft" constructions: *it was noted*, *it was decided to* and *it seemed likely that*.

Referential expressions are usually very close to their referents in the text. In *it had not invaded*, the pronoun refers to the immediately preceding noun phrase *area of tumor*. The pronoun in *it did partially occlude* has the same referent, but in this case there are two intervening nouns: *epithelium* or *trachea*. Thus, a rule that assigns pronouns to the most recent noun would work for the first case, but not for the second.

The syntactic position of a potential referent is an important factor. For example, a referent in subject position is a more likely candidate than the direct object, which in turn is more likely than an object of a preposition. In the fifth sentence of the text above, the pronoun *he* could refer to the patient or to the physician. The proper noun *Dr. Smith* is the more likely candidate, because it is the subject of the preceding sentence.

Centering theory accounts for reference by noting how the center (focus of attention) of each sentence changes across the discourse (Grosz et al. 1995). In our example text, the patient is the center of the first three sentences, the physician is the center of the fourth and fifth sentence, and the area of tumor is the center of the last sentence. In this approach, resolution rules attempt to minimize the number of changes in centers. Thus, in the above text it is preferable to resolve *he* in sentence five as the physician rather than the patient because it results in smoother transition of centers.

Semantic information for resolving referential expressions involves consideration of the semantic type of the expression, and how it relates to potential referents (Hahn et al. 1999):

- semantic type is the same as the potential referent
- semantic type is a subtype or a parent of the potential referent
- semantic type has a close semantic relationship with the potential referent

For example, in the example text, the definite noun phrase *the balloon* must be resolved. If the phrase *a balloon* occurred previously, this would be the most likely referent. Since there is no previous noun of similar type, it is necessary to establish a semantic relationship with a preceding noun. The word *dilation* is the best candidate because a balloon is a medical device used by that procedure.

8.5.6 Evaluation Metrics

Evaluating the performance of an NLP system is critical whether the NLP system is targeted for an end user directly or is part of a larger application. Evaluation generally involves obtaining a **reference standard** with which the system can be compared against. In the clinical domain, creating such reference standards can be a challenge, because of the need for anonymized patient information so the reference standard (also called gold standard) can be shared among researchers at different institutions. When a reference standard is created, it is important to further reliability and to reduce subjectivity and bias by relying on multiple experts, by measuring the inter-rater agreement, and by randomly selecting text instances (Friedman et al. 1998, Hripcsak and Wilcox 2002). There are basically two types of evaluation, **extrinsic** and **intrinsic**.

In an **extrinsic evaluation**, the NLP system is part of a bigger application that is intended to achieve a real-world task, and the aim of the evaluation is to measure performance of the overall task. Therefore, the output of the NLP system is not evaluated independently. A reference standard is generally created using domain experts, who manually perform the task by reading the text, or the reference standard may already exist from another study. For instance, an NLP system may be part of a clinical decision support system aimed at identifying patients with pneumonia based on textual radiology reports. A reference standard would be created consisting of radiology reports that would have been randomly selected and then reviewed manually as denoting or not denoting *pneumonia*. Domain experts are best to use to obtain a reference standard since they routinely read clinical reports and interpret the information in them. A substantial portion of the results of an extrinsic evaluation may be attributable to the decision support component, which generally will include reasoning based on the information extracted by the NLP system. For example, reasoning may be necessary in order to associate extracted findings, such as *infiltrate* and *consolidation*, with *pneumonia*. In such an evaluation, an error analysis would be needed to differentiate NLP errors from errors in the decision support component of the system.

In an **intrinsic evaluation**, the output of the NLP system is evaluated by comparing the output against a reference standard, which generally has been manually annotated so that it is deemed accurate. The extent of differences in results obtained by the NLP system and the reference standard are then computed. Evaluation may be performed for each component of the NLP system, or for the overall results. Depending on the evaluation design and NLP task, annotation generally requires linguistic expertise and may also require domain expertise. Therefore, an intrinsic evaluation that focuses on performance of the NLP system is helpful for advancing NLP development. However, generating the reference standard is generally time-consuming, and may not adequately reflect the information needed for a subsequent real-world application.

There are three basic quantitative measures used to assess performance in an extrinsic or intrinsic evaluation. They are all calculated from the number of **true positives** (TP), **true negatives** (TN), and **false negatives** (FN).

Recall is the percentage of results that should have been obtained according to the test set that actually were obtained by the system:

Recall = Number of correct results obtained by system (TP)/
Number of results specified in gold standard (TP + FN)

Precision is the percent of results that the system obtained that were actually correct according to the test set:

Precision = Number of correct results obtained by system (TP)/
Total number of results obtained by system (TP + FP)

There is usually a tradeoff between recall and precision, with higher precision usually being attainable at the expense of recall, and *vice versa*. The **F measure** is a combination of both measures and can be used to weigh the importance of one measure over the other by giving more weight to one. If both measures are equally important, the F measure is the **harmonic mean** of the two measures.

When reporting results, an **error analysis** provides much insight into ways to improve a system. This process involves determining reasons for errors in recall and in precision. In an extrinsic evaluation, some errors could be due to the NLP system and other errors could be due to the subsequent application component. Some NLP errors in recall (i.e. false negatives) could be due to failure of the NLP system to tokenize the text correctly, to recognize a word, to detect a relevant pattern, or to correctly interpret the meaning of a word or a structure correctly. Some errors in precision could be due to errors in interpreting the meaning of a word or structure or to loss of important information. Errors caused by the application component could be due to failure to access the extracted information properly or failure of the reasoning component.

8.6 Issues for NLP in Biomedicine and Health

Natural language processing is challenging for general language, but there are issues that are particularly germane to the domains of biomedicine and health. We list a few of them in this section.

8.6.1 Patient Privacy and Ethical Concerns

As an NLP system deals with patient information, its designers must remain cognizant of the privacy and ethical concerns entailed in handling protected health information. In the clinical domain for instance, the **Health Insurance Portability and Accountabiliy Act** (HIPAA; see Chap. 10) regulates the protection of patient-sensitive information (see Chap. 10 for a detailed description of privacy matters in the clinical domain). Online, patients provide much information about their own health in blogs and online communities. While there are no regulations in place concerning online patient-provided information, researchers have established guidelines for the ethical study and processing of patient-generated speech (Eysenbach and Till 2001).

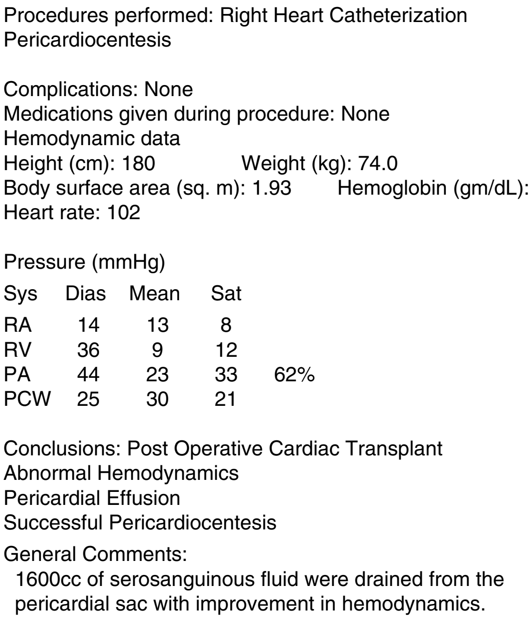
8.6.2 Good System Performance

If the output of an NLP system is to be used to help manage and improve the quality of health care and to facilitate research, it must have high enough performance for the intended application. Evidently different applications require varying levels of performance; however, the performance must generally not be significantly worse than that of domain experts. This requirement means that before an NLP-based system can be used for a practical task, it must be evaluated carefully, both intrinsically and extrinsically, in the setting where the system will be used. For instance, if a system is designed for clinicians in the ICU, testing its use with primary care physicians might not be a reliable evaluation. This point is valid for the biological and health consumer domains as well (Caporaso et al. 2008).

8.6.3 System Interoperability

NLP-based systems are often part of larger applications. There must be seamless integration of the NLP component into its parent application. This point is valid for any domain, but becomes particularly relevant in the clinical domain, where NLP is typically part of the electronic record. In practice, the following might have to be ensured, depending on the particular task of the application:

- The system has to follow standards for interoperability among different health information technology systems, such as **Health Level 7** (**HL7**) and the **Clinical Document Architecture** (CDA; see Chap. 7). This is particularly important for information that pertains to a patient, but is not part of the notes per se. Similarly, the system has to be aware of the controlled terminologies in use in the institution (e.g. **SNOMED-CT** and **ICD-9-CM**; see Chap. 7), so that its input and output are understood by the clinical information system.
- The system has to be aware of the information storage strategies of the clinical information system. As of this date, there is no established structure across different institutions, and so care has to be given to understanding the structure of a particular institution in which the NLP system will be deployed. For instance, it is possible that different note types and reports are stored in different ways in the clinical information system, and the NLP system will have to acknowledge this in its workflow. Furthermore, within a given note, there might be institution-specific conventions concerning the overall format and use of abbreviations. Figure 8.8 shows portions of a cardiac catherization report. Some of the sections contain free text (i.e. procedures performed, comments, general conclusions), some consist of structured fields (i.e. height, weight) that are separated from each other by white space, and some consist of tabular data (i.e. blood pressure). The NLP system has to be able to recognize and handle these different formats.



Procedures performed: Right Heart Catheterization
Pericardiocentesis

Complications: None
Medications given during procedure: None
Hemodynamic data
Height (cm): 180 Weight (kg): 74.0
Body surface area (sq. m): 1.93 Hemoglobin (gm/dL):
Heart rate: 102

Pressure (mmHg)

Sys	Dias	Mean	Sat	
RA	14	13	8	
RV	36	9	12	
PA	44	23	33	62%
PCW	25	30	21	

Conclusions: Post Operative Cardiac Transplant
Abnormal Hemodynamics
Pericardial Effusion
Successful Pericardiocentesis

General Comments:
1600cc of serosanguinous fluid were drained from the pericardial sac with improvement in hemodynamics.

Fig. 8.8 A portion of a sample cardiac catherization report

- The system has to generate output that can be stored in a **clinical data repository** (CDR) or **clinical data warehouse** (CDW). This is especially true for NLP systems that are deployed as part of an operational clinical application or research application. As for clinical information systems, there is no established database schema for the type of rich information found in clinical records. Depending on the type of NLP (i.e. information extraction vs. full syntactic parsing vs. semantic parsing), there might be some loss of information when storing the output of the NLP system in a database.

8.6.4 Misspellings and Typographical Errors

Clinicians, when typing free-text in the EHR, do so under time pressure and generally do not have the time to proofread their notes carefully.

In addition, they frequently use abbreviations (e.g. *HF* for *Hispanic female* or *heart failure*, *2/2* for *secondary to* or a date), many of which are non-standard and ambiguous. For patients and health consumers, when posting content online, misspellings, typographical errors, and non-standard abbreviations are pervasive like in the rest of the social Web. In a breast cancer online community, for instance, we found frequent spelling variations for the drug *tamoxifen* (e.g. *tamoxifin* and *tamaxifen*) as well as abbreviations. Ignoring these variations may cause an NLP system to lose or misinterpret information. At the same time, errors can be introduced when correcting the typos automatically. For instance, it is not trivial to correct *hyprtension* automatically without additional knowledge because it may refer to *hypertension* or *hypotension*. This type of error is troublesome not only for automated systems, but also for clinicians when reading a note, as this phenomenon is aggravated by the large amount of short, misspelled words in notes.

8.6.5 Expressiveness Vs. Ease of Access

Natural language is very expressive. There are often several ways to express a particular medical concept as well as numerous ways to express modifiers for that concept. For example, ways to express severity include *faint*, *mild*, *borderline*, *1+*, *3rd degree*, *severe*, *extensive*, *mild*, and *moderate*. This expressive power makes it challenging to build information extraction systems that capture all modifiers of a concept with high recall. Often, to complicate matters, modifiers can be composed or nested. For instance in the phrase "*no improvement in pneumonia*," *improvement* is a change modifier that modifies the concept *pneumonia*, and *no* is a negation marker that modifies *improvement* (not *pneumonia*). In this situation, an information extraction system that detects changes concerned with pneumonia would have to look for primary findings associated with pneumonia, filter out cases not associated with a current episode, look for a change modifier of the finding, and, if there is one, make sure there is no negation modifier on the change modifier. An alternative representation would facilitate retrieval by flattening the nesting. In this case, some information may be lost but ideally only information that is not critical. For example, *slightly improved* may not be clinically different from *improved* depending on the application. Since this type of information is fuzzy and imprecise, the loss of information may not be significant. However, the loss of a negation modifier would be significant, and those cases should be handled specially. Another such example concerns hedging, which frequently occurs in radiology reports as well as in scientific articles. This tradeoff between capturing the full expressive power of natural language and ease of access when extracting information influences design choices for an NLP system and depends on the overall task for which the system is built.

8.6.6 Reliance on Medical Knowledge and Reasoning

Whether in biomedical, clinical or health consumer texts, there is much implicit knowledge present in the text. In some systems, recovering the missing information can be important. For instance, the phrase "I have a temperature" as written by a patient online can mean I have a fever, but "I have a temperature of 98.6" means no fever. Inferring the presence of fever from the presence and/or value of a numerical modifier requires external medical knowledge. Similarly in the clinical domain, interpreting the findings extracted from a chest radiological report, or inferring that a patient is depressed based on the fact that an anti-depressant is prescribed (even though there is no explicit mention of depression in a note) requires extensive medical knowledge. Such knowledge can be quite complex. Ontologies contain some of this knowledge, encoded through entities and relations (e.g. parent–child, part of) between the entities. But the ontology may not be complete enough for particular tasks, both in the coverage of entities and relations, and such knowledge would have to be acquired. One way to do so is to leverage knowledge from experts and encode manual rules. For instance, a rule that detects a comorbidity of neoplastic disease based on information in a discharge summary, could consist of a Boolean combination of over 200 terms (Chuang et al. 2002). Another approach is to leverage machine-learning techniques and learn rules from examples of texts. In a supervised framework, for such rules to be learned, a large number of training and testing instances must be annotated. The cost (both from a financial and time standpoint) to annotate these instances depends on many factors, including the difficulty of the task and level of medical expertise needed to carry out accurate annotation, the degree of subjectivity entailed in the task, the number of instances needed to annotated and the easy access to annotators. Whether medical knowledge should be encoded directly or learned through examples is a question that is beyond the focus of this chapter. Thus far, the most promising approaches in NLP are hybrid ones that combine existing medical knowledge with knowledge mined from text.

8.6.7 Domains and Subdomains

As discussed earlier in this chapter, the fact that texts belong to a particular domain, be it clinical, biological or related to health consumers, allows us to capture domain-specific characteristics in the lexicon, the grammar, and the discourse structure. Thus, the more specific the domain of a text, the more knowledge can be encoded to help its processing, but the NLP system would then be very limited and specialized. For instance, in the domain of online patient discourse, patients discussing breast cancer among their peers online rely on a very different set of terms than caregivers of children on the autism spectrum. One could develop a lexicon for each subdomain, e.g. online breast cancer patients and online autism caregivers. But maintaining separate lexicons can be inefficient and error prone, since there can be a significant amount of overlap among terms across subdomains. Conversely, if a single lexicon is developed for all subdomains, ambiguity can increase as terms can have different meanings in different subdomains. For example, in the emergency medicine domain *shock* will more likely refer to a procedure used for resuscitating a patient, or to a critical condition brought about by a drop in blood flow, whereas in psychiatry notes it will more likely denote an emotional response or electric shock therapy. Deciding on whether to model a domain as a whole or to focus on its subdomains independently of each other is a tradeoff. Careful determination of the use cases of a system can help determine the best choice for the system.

8.6.7.1 Dynamic Nature of Biomedical and Health Domains

Change is a natural phenomenon of human language. With time, new concepts enter the language and obsolete terms fall out of use. The biomedical and health domains are highly dynamic in the influx of new terms (e.g. new drug names, but also sometimes new disease names, like SARS and H1N1). The biomolecular domain is particularly dynamic. For example, for the week ending July 20, 2003, the Mouse Genome Informatics Website (Blake et al. 2003) reported 104 name changes related to the mouse alone. If the other organisms being actively sequenced were also considered, the number of name changes during that week would be much larger. Related to the language change is the sheer number of entities relevant to the biomolecular domain. Table 8.4 shows the approximate numbers for some different types of entities in the biomolecular domain. The number of entities is actually larger because not all types are shown in the table (i.e. small molecules, cell lines, genes and proteins of all organisms). Having such a large number of names means the NLP system must have a very large knowledge base of names or be capable of dynamically recognizing the type by considering the context. When entities are dynamically recognized without use of a knowledge source, identifying them within an established nomenclature system is not possible.

Table 8.4 The approximate number of some types of biomolecular entities

Type of entity	Number
Gene	3.5×10^4 (human only)
Protein	$>10^5$ (human only)
Cell type	10^6
Species	10^7

8.6.8 Polysemy

Biomedical and health terms are often ambiguous. This is particularly pervasive in the biomolecular domain because of the high number of acronyms and abbreviations. Short symbols consisting of two to three letters are frequently used that correspond to names of biomolecular entities. Since the number of different combinations consisting of only a few letters is relatively small, individual names often correspond to different meanings. For example, *to*, a very frequent English word, corresponds to two different Drosophila genes and to the mouse gene *tryptophan 2,3-dioxygenase*. *To further complicate matters*, the names of genes in different model organism groups are established independently of each other, leading to names that are the same but which represent different entities. The ambiguity problem is actually worse if the entire domain is considered. For example, *cad* represents over 11 different biomolecular entities in Drosophila and the mouse but it also represents the clinical concept *coronary artery disease*. Another contributing factor to the ambiguity problem is due to the different naming conventions for the organisms. These conventions were not developed for NLP purposes but for consistency within individual databases. For example, Flybase states that "Gene names must be concise. They should allude to the gene's function, mutant phenotype or other relevant characteristic. The name must be unique and not have been used previously for a Drosophila gene." This rule is fairly loose and leads to ambiguities.

8.6.9 Synonymy

Complementary to the phenomenon of polysemy, there are often terms that are different variations of the same concepts. For instance, the term *blood sugar* is often used by health consumers to refer to a *glucose measurement*, but it is used rarely if ever in the clinical literature or in clinical notes. Synonymy occurs across domains, but is also present within a single domain. In the biomolecular domain, for instance, names are created within the model organism database communities, but they are not necessarily exactly the same as the names used by authors when writing journal articles. There are many ways authors vary names (particularly long names), which leads to difficulties in named entity recognition. This is also true in the medical domain but the problem is exacerbated in the biomolecular domain because of the frequent use of punctuations and other special types of symbols. Some of the more common types of variations are due to punctuations and use of blanks (*bmp-4*, *bmp 4*, *bmp4*), numerical variations (*syt4*, *syt IV*), variations containing Greek letters (*iga*, *ig alpha*), and word order differences (*phosphatidylinositol 3-kinase*, *catalytic*, *alpha polypeptide*, *catalytic alpha polypeptide phosphatidylinositol 3-kinase*). Low-level text processing can resolve some of these types of synonymy.

8.6.10 Complexity of Biological Language

The semantic interpretation of biological language is complex. In clinical text, the important source of information typically occurs in noun phrases that consist of descriptive information that corresponds to named entities and their modifiers. In biomolecular text, important information often involves interactions that are highly nested, corresponding to verb phrases, which are more complex structures than noun phrases and are often highly nested. Although syntactically the interaction may occur as a noun, it is generally a nominalized verb, and thus, has arguments that are important to capture along with their order (e.g. *Raf-1 activates Mek-1* has a different meaning than *Mek-1 activates Raf-1*). In addition, the argument may also be another interaction. Thus a typical sentence usually contains several nested interactions. For example, the sentence "*Bad phosphorylation induced by interleukin-3* (*il-3*) *was inhibited by specific inhibitors of phosphoinositide 3-kinase*

Table 8.5 *Nested interactions extracted from the sentence* Bad phosphorylation induced by interleukin-3 (il-3) was inhibited by specific inhibitors of phosphoinositide 3-kinase (pi 3-kinase)

Interaction	Argument 1 (agent)	Argument 2 (target)	Interaction id
Phosphorylate	?	Bad	1
Induce	Interleukin-1	1	2
Inhibit	?	Phosphoinositide 3-kinase	3
Inhibit	3	1	

A '?' denotes that the argument was not present in the sentence

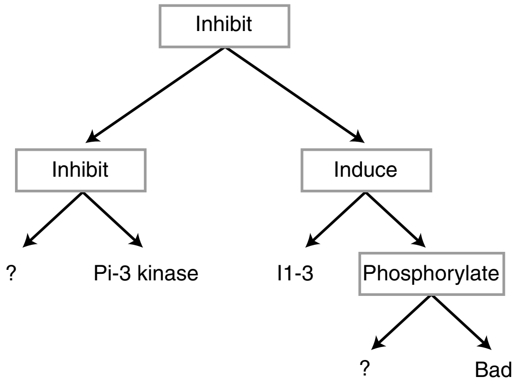


Fig. 8.9 A tree showing the nesting of biomolecular interactions that are in the sentence "*Bad phosphorylation induced by interleukin-3* (*il-3*) *was inhibited by specific inhibitors of phosphoinositide 3-kinase* (*pi 3-kinase*)"

(*pi 3-kinase*)" consists of four interactions (and also two parenthesized expressions specifying abbreviated forms). The interaction and the arguments are illustrated in Table 8.5. The nested relations can be illustrated more clearly as a tree (see Fig. 8.9). Notice that the arguments of some interactions are also interactions (i.e. the second argument of *induce* is *phosphorylate*). Also note that an argument that is not specified in the sentence is represented by a "?" in the figure.

8.6.11 Interactions among Linguistic Levels

While the earlier sections of this chapter have introduced each linguistic level on its own, it should be clear through the many examples presented throughout the chapter that processing of language is not as simple as applying a pipeline of independent modules- one to determine tokens, one to assign part-of-speech tags to tokens and to parse the syntax, one to interpret the meaning of a sentence, one to resolve the discourse-level characteristics of the text, and so on. In reality, all linguistic levels influence each other. Low-level decisions about how to tokenize a string impact named-entity recognition; determining which sense to attribute to a named entity depends on its place in the syntactic tree, the pragmatics of the text, and its place in the discourse structure. Determining how to model these interactions is one of the primary open research questions of natural language processing.

8.7 Resources for NLP in Biomedicine and Health

One of the ways for the field to make progress is for different teams of researchers to share their datasets, tools and resources. Shared datasets allow different research teams to test and compare their systems on the same data. Annotated shared datasets are critical, as they allow teams to train their systems as well. As such, they are very valuable to the community. In recent years, there has been a strong push in the biological and clinical NLP communities to create publicly available resources and tools, and to conduct community challenges. We present a few of them here but, because of the explosion of resources in the field, this list is bound to be obsolete. We therefore encourage the reader to check the literature and the Web for the latest.

8.7.1 Databases and Lexicons

- UMLS (Including the Metathesaurus, Semantic Network, the Specialist Lexicon) – Can be used as a knowledge base and resource for a lexicon. The Specialist lexicon provides detailed syntactic knowledge for words and phrases, and includes a comprehensive medical vocabulary. It also provides a set of tools to assist in NLP, such as a lexical variant generator, an index of words corresponding to UMLS terms, a file of derivational variants (e.g.

abdominal, *abdomen*), spelling variants (e.g. *fetal*, *foetal*), and a set of neoclassical forms (e.g. *heart*, *cardio*). The UMLS Metathesaurus provides the concept identifiers, while the Semantic Network specifies the semantic categories for the concepts. The UMLS also contains the terminology associated with various languages (e.g. French, German, Russian). The UMLS is the union of several vocabularies that are particularly useful for NLP, such as SNOMED-CT, LOINC, and MeSH.

- MedDRA and RxNorm are terminologies specific to adverse event terminology and medications. They are particularly helpful in the clinical domain, in pharmacovigilance and in pharmacogenomics.
- Biological databases. These include Model Organism Databases, such as Mouse Genome Informatics (Blake et al. 2003), the Flybase Database (FlyBase Consortium 2003), the WormBase Database (Harris et al. 2003), and the *Saccharomyces* Database (Issel-Tarver et al. 2001), as well as more general databases GenBank (Benson et al. 2003), Swiss-Prot (Boeckmann et al. 2003), LocusLink (Pruitt and Maglott 2001), and the Gene Ontology (GO 2003).

8.7.2 Corpora

- PubMed Central[2] provides full-text articles in biomedicine and health. PubMED provides abstracts and useful meta-information, such as MeSH indexes, and journal types.
- The MIMIC II database collects de-identified data about patients in the intensive care unit (Saeed et al. 2002). Along with ICU-specific structured data and times series, there are nursing notes, progress notes and reports available.
- The Pittsburgh Note Repository[3] provides de-identified clinical notes of many different types. Some of the notes have been annotated with specific semantic information as part of community challenges.

8.7.3 Community Challenges and Annotated Corpora

In biology, the GENIA corpus contains articles annotated with syntactic, semantic and discourse information (Kim et al. 2003). The BioCreAtIvE challenges have provided annotated articles with biologically relevant named-entities and entity-fact associations, such as protein-functional term association (Hirschman et al. 2005). The BioScope corpus provides texts annotated with hedging and negation information (Vincze et al. 2008). As part of yearly community challenges in the clinical domain, there are several corpora currently available with different types of annotations. The earliest is a collection of radiology reports with ICD-9-CM codes (Pestian et al. 2007). The i2b2 community challenges have provided clinical notes annotated with smoking status (Uzuner et al. 2008), obesity and co-morbidities (Uzuner 2009), medications mentions (Uzuner et al. 2010), assertions (Uzuner et al. 2011), and more recently co-references. There are several corpora specific to word sense disambiguation (WSD) for different semantic classes. The National Library of Medicine provides an annotated WSD dataset for 50 frequently occurring ambiguous terms based on the 1998 version of MEDLINE (Weeber et al. 2001).

8.7.3.1 Annotation Schema

In the same way terminologies like the UMLS provide an established organization for concepts in the language, community efforts have just started to create established representations for certain aspects of information, such as the different modifiers of concepts and the relations among concepts that can occur in texts. There had been early efforts by a large number of researchers called *The Canon group* to create such standard (Evans et al. 1994). That effort resulted in a common model for radiological reports of the chest (Friedman et al. 1995), but the model was not actually utilized by the different researchers.

8.7.3.2 Tools

Since the NLP field is currently a very active area of research, and new tools are continually

[2] http://www.ncbi.nlm.nih.gov/pmc (Accessed 4/26/13).

[3] http://www.dbmi.pitt.edu/nlpfront. (Accessed 4/26/13).

being developed; we point the reader to ORBIT (Online Registry of Biomedical Informatics Tools[4]). It provides a repository of tools, maintained by the community, and is a good place to get access to the most recent tools. In the general NLP domain, there are a few valuable suites of tools available, including NLTK,[5] LingPipe,[6] and OpenNLP.[7] Finally, UIMA[8] is a general framework for text analysis that is gaining popularity in NLP.

Suggested Readings

Harris, Z., Gottfried, M., Ryckmann, T., Mattick, P., Jr., Daladier, A., Harris, T. N., & Harris, S. (1989). *The form of information in science: Analysis of an immunology sublanguage*. Reidel/Dordrecht: Boston Studies in the Philosophy of Science. This book offers an in-depth description of methods for analyzing the languages of biomedical science. It provides detailed descriptions of linguistic structures found in science writing and the mapping of the information to a compact formal representation. The book includes an extensive analysis of 14 full-length research articles from the field of immunology, in English and in French.

Sager, N., Friedman, C., & Lyman, M. S. (1987). *Medical language processing: Computer management of narrative data*. New York: Addison-Wesley. This book describes early techniques used by the Linguistic String Project, a pioneering language processing effort in the biomedical field, explaining how biomedical text can be automatically analyzed and the relevant content summarized.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing. An introduction to natural language processing, computational linguistics and speech recognition*. Upper Saddle River: Prentice Hall.

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press. NLP is a very active field of research in the general domain. Many of the applications and techniques described in this chapter are investigated in other domains. For a review of NLP methods in the general domain, we refer the reader to the above three textbooks.

Kübler, S., McDonald, R., & Nivre, J. (2009). *Dependency parsing. Synthesis lecture on human language technology*. Morgan & Claypool.This book provides an in-depth review of dependency parsing in the general domain.

Palmer, M., Gilder, D., & Xue, N. (2010). *Semantic role labeling. Synthesis lectures in human language technology*. Morgan & Claypool.This book provides an in-depth discussion of semantic parsing.

Questions for Discussion

1. Develop a regular expression to regularize the tokens in lines four to nine of the cardiac catheterization report shown in Fig. 8.8 (*Complications* through *Heart Rate*).
2. Create a lexicon for the last seven lines of the cardiac catheterization report shown in Fig. 8.8 (*Conclusions* through the last sentence). For each word, determine all the parts of speech that apply, using the tags in Table 8.1. Which words have more than one part of speech? Choose eight clinically relevant words in that section of the report, and suggest appropriate semantic categories for them that would be consistent with the SNOMED-CT terminology and with the UMLS semantic network.
3. Using the grammar in Fig. 8.3, draw a parse tree for the last sentence of cardiac catheterization report shown in Fig. 8.8.
4. Using the grammar in Fig. 8.4, draw parse trees for the following sentences: *no increase in temperature*; *low grade fever*; *marked improvement in pain*; *not breathing*. (Hint: some lexemes have more than one word.)
5. Identify all the referential expressions in the text below and determine the correct referent for each. Assume that the compute attempts to identify referents by finding the most recent noun phrase. How well does this resolution rule work? Suggest a more effective rule.

[4] orbit.nlm.nih.gov (Accessed 4/19/13).

[5] www.nltk.org (Accessed 4/18/13).

[6] www.alias-i.com/lingpipe/ (Accessed 4/18/13).

[7] http://incubator.apache.org/opennlp/ (Accessed 4/19/13).

[8] http://uima.apache.org/index.html (Accessed 4/19/13).

The patient went to receive the AV fistula on December 4. However, he refuses transfusion. In the operating room it was determined upon initial incision that there was too much edema to successfully complete the operation and the incision was closed with staples. It was well tolerated by the patient.

6. In the two following scenarios, an out-of-the-shelf NLP system that identifies terms and normalizes them against UMLS concepts, is applied to a large corpus of texts. In the first scenario, the corpus consists of patient notes. Looking at the frequency of different concepts, you notice that there is a large number of patients with the concept C0019682 (HIV) present, much larger than the regular incidence of HIV in the population reported in the literature. In the second scenario, the corpus consists of full-text biology articles published in PubMED Central. Looking at the frequency of different concepts, you notice that the failed axon connection (fax) gene is one of the most frequently mentioned genes in your corpus. Describe how you would check the validity of these results. For both cases, discuss what could explain the high frequency counts.