CS 5316–Natural Language Processing
Assignment 1
Deadline: Jan. 31 (Wednesday) at 11:55 PM

Question 1
Provide a definition and some examples of the following branches of linguistics:

a. Phonetics
Phonetics is the study of speech sounds as they stand in isolation. Simply put, **it** is spelling words the way they sound. While each letter in English is assigned at least one sound, there are lots of letters or letter combinations that are pronounced differently in different words. Example: the letter 'p' is frequently assigned the /p/ as in the word 'paper,' however, if it's paired with the letter 'h,' as in 'phone,' the 'ph' together make the /f/.

b. Phonology
Phonology is the study of speech sounds and how they change depending on certain situations or placements in syllables, words, and sentences. Example: the final 's' sounds in 'helps' and 'crabs' follow a simple-to-understand phonological rule. In these words, the 's' sound changes depending on what speech sound immediately precedes it. The ending in ‘helps’ sounds closer to ‘s’, however that in ‘crabs’ sounds more like a ‘z’.

c. Morphology
Morphology is the arrangement and relationships of the smallest meaningful units in a language. In order to understand morphology, you need to know the term morpheme, which is the smallest unit of a word with meaning. When a number of letters are put together into a word part that now has meaning, then you have a morpheme. Morphology studies how these units of meaning, or word parts, can be arranged in a language. Example: “Firehouse” begins with an -f sound and ends with the -s sound. However, those sounds alone don't have meaning. Fire means bright light while house means a dwelling for human beings. Putting these together creates a completely new word.

d. Syntax
Syntax is the part of linguistics that studies the structure and formation of sentences. Syntax deals with phrase and sentence formation out of words. A sentence could make no sense and still be correct from the syntax point of view as long as words are in their appropriate spots and agree with each other. Example: ''Colorless green ideas sleep furiously.'' To create grammatically correct and acceptable English sentences, we have to follow the English rules for syntax.

e. Semantics
Semantics means the meaning and interpretation of words, signs, and sentence structure. Semantics largely determine our reading comprehension, how we understand others, and even what decisions we make as a result of our interpretations. Semantics can also refer to the branch of study within linguistics that deals with language and how we understand meaning. Example: “She's drowning in a sea of grief.” – To make sense this has to be taken figuratively, and not literally.

f. Pragmatics
Pragmatics is a branch of linguistics, which is the study of language. Pragmatics focuses on conversational implicature, which is a process in which the speaker implies and a listener infers. Simply put, pragmatics studies language that is not

directly spoken. Instead, the speaker hints at or suggests a meaning, and the listener assumes the correct intention. In a sense, pragmatics is seen as an understanding between people to obey certain rules of interaction. Example: “How are you today?” And instead of responding with your health problems, family issues you simply say something in the lines of “Fine, how are you?”

Question 2

Answer the following

a. Define grammar. Can natural language (English) grammar be explicitly coded?
   Grammar is the whole system and structure of a language or of languages in general, usually taken as consisting of syntax and morphology (including inflections) and sometimes also phonology and semantics.
   No, natural language grammar cannot be explicitly coded because the way people actually speak English does not necessarily follow the grammatical rules. A lot of idioms and slang exist in natural language, which is impossible for a computer to interpret as they do not have a definite structure.
b. What are the common parts-of-speech in English?
   1) Noun This part of a speech refers to words that are used to name persons, things, animals, places, ideas, or events.
      e.g. Tom Hanks, Dogs
   2) Pronoun A pronoun is a part of a speech which functions as a replacement for a noun. Some examples of pronouns are: I, it, he, she, mine, his, hers, we, they, theirs, and ours.
      e.g. She, Mine, We
   3) Adjective This part of a speech is used to describe a noun or a pronoun. Adjectives can specify the quality, the size, and the number of nouns or pronouns.
      e.g. intricate, two, huge
   4) Verb This is the most important part of a speech, for without a verb, a sentence would not exist. Simply put, this is a word that shows an action (physical or mental) or state of being of the subject in a sentence.
      e.g. missed, are, am
   5) Adverb Just like adjectives, adverbs are also used to describe words, but the difference is that adverbs describe adjectives, verbs, or another adverb.
      e.g. gracefully, everywhere
   6) Preposition This part of a speech basically refers to words that specify location or a location in time.
      e.g. above, below, throughout
   7) Conjunction The conjunction is a part of a speech which joins words, phrases, or clauses together.
      e.g. and, yet, but
   8) Interjection This part of a speech refers to words which express emotions. Since interjections are commonly used to convey strong emotions, they are usually followed by an exclamation point.
      e.g. Ouch, Hurray, Hey
c. Distinguish between a phrase and clause.
   A clause can be completely meaningful in which case it becomes “main” clause or partially meaningful and depends on the other clause for its meaning in that case it becomes “subordinate” clause.

A phrase is just a collection of words which can or cannot have a literal meaning given by the words in the phrase.
e.g Pull over, Pull off, Take off
e.g for Clause: so that I could give him the book, I ran faster than possible.

d. What is transliteration? Distinguish between forward and backward transliteration.
Transliteration is converting the text from one script to another. It does not render meaning. This means changing only the source letters or characters into corresponding those of the target language.
For example (from Greek to English):-
Ελληνική Δημοκρατία = Hellenic Republic (translation)
Ελληνική Δημοκρατία = Ellēnikḗ Dēmokratía (transliteration)

Forward transliteration, mostly used for translating personal and location names, is a way of translating source names into target language with approximate phonetic equivalents, while backward transliteration traces back to the foreign names.

Question 3

a. Read about IBM Watson and its win in Jeopardy. Summarize your findings in a paragraph.
Watson is an IBM supercomputer that combines artificial intelligence (AI) and sophisticated analytical software for optimal performance as a "question answering" machine. Watson was named after IBM's first CEO, industrialist Thomas J. Watson. The computer system was specifically developed to answer questions on the quiz show Jeopardy. Initially it played 100 games against past winners in an effort to improve his chances of winning. While in 2011, the Watson computer system competed on Jeopardy against former winners Brad Rutter and Ken Jennings and won the game with $77,147 leaving Rutter and Jennings in the dust with $21,600 and $24,000 respectively. The victory of IBM's Watson over two human contestants was the first, and possibly only, time the machine impressed itself on the general public's consciousness.

b. Study IBM Watson as a service / product of IBM. List down its key features, and provide an overview of its key techniques.
Watson is a question answering computer system capable of answering questions posed in natural language. IBM Watson is a system based on cognitive computing. Cognitive computing is a technique which is a mixture of different techniques such as machine learning, natural language processing, artificial intelligence, human interaction, reasoning etc.
So, overall IBM Watson is made up of these techniques. In short, IBM Watson can think like us, learn like us and yes, it can give answers to our questions.
Software:
Watson uses IBM's DeepQA software and the Apache UIMA framework. The system was written in various languages, including Java, C++, and Prolog, and runs on the SUSE Linux Enterprise Server 11 operating system using the Apache Hadoop framework to provide distributed computing.
Hardware:
The system is workload-optimized, integrating massively parallel POWER7 processors and built on IBM's DeepQA technology. Watson employs a cluster of ninety IBM Power 750 servers, each of which uses a 3.5 GHz

POWER7 eight-core processor, with four threads per core. In total, the system has 2,880 POWER7 processor threads and 16 terabytes of RAM

Question 4

Study NLTK and spaCy, toolkits on NLP. Describe how textual data can read and tokenized using these toolkits.

In order to tokenize the text, the simplest approach used by NLTK is to split on whitespace. Some of the functions used by NLTK for tokenization are: sent_tokenize() – used for separating out each sentence in a text word_tokenize() – used for separating out words but might contain punctuations as well e.g. commas, periods, etc. wordpunct_tokenize() – used for correctly separating out each symbol and word separately.

Incase of spaCy, data in the form of raw text can easily be read with nlp function and returned a Doc which is then toekenized. Each tokenizer consults a mapping table TOKENIZER_EXCEPTIONS, which allows sequences of characters to be mapped to multiple tokens. Each token may be assigned a part of speech and one or more morphological features. After tokenization, spaCy can parse and tag a given Doc.