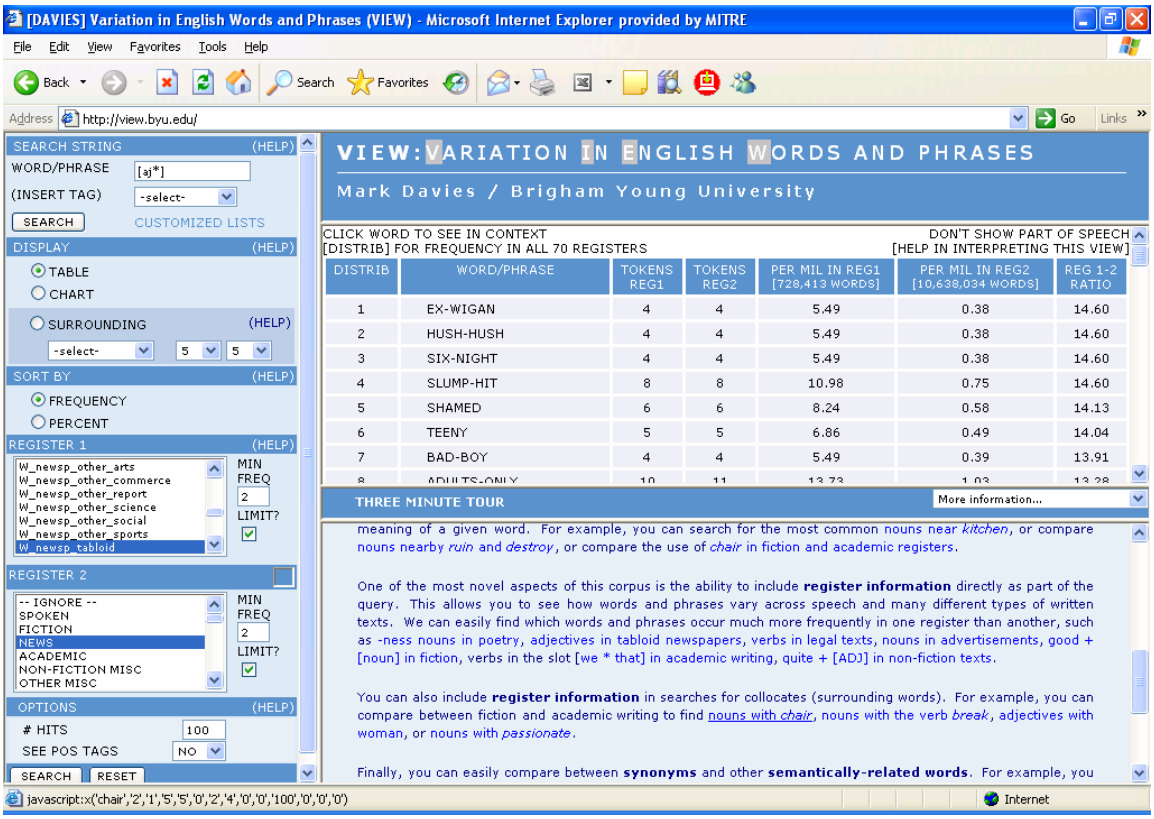
Computational linguistics - Inderjeet Mani

1. Consider the sentence “Time flies.” What ambiguities does the sentence present to a computer, and how might these be resolved?

2. Professor Oliver Smart-Botham has claimed that tabloid ‘newspapers’ use very different language from ordinary newspapers, approximating an ‘argot-like sublanguage designed to grab the reader’s attention with sensational revelations’. Investigate whether this claim is borne out in the case of adjectives in the British National Corpus (BNC), a 100 million-word text corpus capturing written and spoken British English. Please use the BNC interface at http://view.byu.edu/.



3. Go the BNC web site at http://www.natcorp.ox.ac.uk/. Find out how often “time” occurs as a verb in the BNC, and how often it occurs as a noun. Do the same for “flies”.

4. Here is some data from Swahili, an agglutinative language:

Swahili sentence	**English translation**
nitakupenda	I will like you
aliwapenda	s/he liked them
atawapenda	s/he will like them

List the Swahili pronouns that translates the pronouns "s/he", "I", "you", and "them". Also list the Swahili auxiliaries that translate past tense and future tense. Finally, identify the Swahili verb that translates the verb "like".

5. Given the table in Problem 4, write a context-free grammar for Swahili sentences. Show the parse tree for *nitakupenda*.

6. Consider the following Spanish grammar:

NP → Nom
Nom → Nom A
Nom → N

N → *computadoras*
N → *hombres*
N → *tacos*
A → *rapidas*
A → *grandes*

Can you list all the expressions that the grammar will generate. Are any ungrammatical? If so, how would you repair the grammar?

7. In discussing pattern-action rules, in particular, example (2) on page 469, we described a regular expression pattern for an individual word: *{V|C}$^+$ {C}ied*. This exercise will involve writing a regular expression pattern for a sequence of words. Just as V stands for any vowel in the set of vowels for a language, and C for any consonant in the set of consonants for a language, we can make N stand for any noun in a set of nouns in the language, A for any adjective, and D for any determiner. Let the set of nouns be *dog*, *cat*, *rat*, the set of adjectives be *tasty*, *thirsty*, *hungry*, and the set of determiners be *the*. Write a regular expression pattern for English NPs. Show some NPs matching the pattern.

Note: First, let's simplify this notation. Instead of using curly parentheses, let's use ordinary ones. So, the above pattern will be *(V|C)$^+$(C)ied*. Second, if there's only one element in a group, we can drop the parentheses. We now have *(V|C)$^+$Cied* . However, you will still need some additional notation. While "X$^+$" stands for the sequence X repeated one or more times, as in the above pattern, "X?" stands for the sequence X repeated zero times (i.e., it doesn't occur) or once, i.e., X is optional.

8. Let us now extend the pattern in problem 7 to accommodate sentences. Let V stand for any verb, R for any adverb, P for any preposition. Let the set of verbs be *eat, drink, run*, and the set of adverbs be *quickly, noisily, hungrily*. Write a regular expression pattern for English sentences. Show some NPs matching the pattern.

Additional notation: Remember, "|" stands for "or". You can group complex patterns needing to be 'or-ed' together with parentheses.

9. Extend the pattern in problem 8 to accommodate sentences containing prepositional phrases, such as *at night the president of the university drinks whisky alone in the house on the lake*.

Additional notation: While "X^+" stands for the sequence X repeated one or more times, as in the pattern in example (2) on page 469, "X*" stands for the sequence X repeated zero or more times.

10. A language called Mukdup, unlike English, lacks determiners, relative pronouns, and tense markers, and is verb-final; like English, however, it has relative clauses succeeding the nouns they modify. For the translation of English *man*, Mukdup uses *mamuo*; for *maniac*, Mukdup uses *mamamuo*; for *is*, Mukdup has *gekwanzo*, for elephant, *koorajo*, and for *shoot*, Mukdup has *boohoomaro*. The suffix *–ni* is used to indicate accusative case for the object of a transitive verb.

Based on this information, and using syntactic categories drawn from the grammar for exercise 14.3 on page 490, draw the parse tree for the Mukdup translation of *The man who shot your elephant is a maniac*.

11. In English, I say "I'm hungry." In German, it's "Ich habe Hunger". In French, it's "J'ai faim" (where "J'" is a contraction for "I", "ai" means "have 1st person present singular" and "faim" means "hunger"). What sort of problems could this pose for a Machine Translation (MT) system? Consider both a direct, a transfer, and an interlingual approach.