IWS 6-7. CORPUS-BASED WORK.

4.5 Exercises

Exercise 4.1 [★★]

As discussed in the text, it seems that for most purposes, we d want to treat some hyphenated things as words (for instance, *co-worker, Asian-American),* but not others (for instance, *ain t-it-great-to-be-a-Texan, child-as-required-yuppie-possession).* Find hyphenated forms in a corpus and suggest some basis for which forms we would want to treat as words and which we would not. What are the reasons for your decision? (Different choices may be appropriate for different needs.) Suggest some methods to identify hyphenated sequences that should be broken up – e.g., ones that only appear as non-final elements of compound nouns:

[$_N$[child-as-required-yuppie-possession] syndrome]

Exercise 4.2 [★★ For linguists]

Take some linguistic problem that you are interested in (non-constituent coordination, ellipsis, idioms, heavy NP shift, pied-piping, verb class alternations, etc.). Could one hope to find useful data pertaining to this problem in a general corpus? Why or why not? If you think it might be possible, is there a reasonable way to search for examples of the phenomenon in either a raw corpus or one that shows syntactic structures? If the answer to both these questions is yes, then look for examples in a corpus and report on anything interesting that you find.

Exercise 4.3 [★★]

Develop a sentence boundary detection algorithm. Evaluate how successful it is. (In the construction of the *Wall Street Journal* section of the ACL-DCI CD-ROM (Church and Liberman 1991), a rather simplistic sentence boundary detection algorithm was used, and the results were not hand corrected, so many errors remain. If this corpus is available to you, you may want to compare your results with the sentence boundaries marked in the corpus. With luck, you should be able to write a system that performs considerably better!)