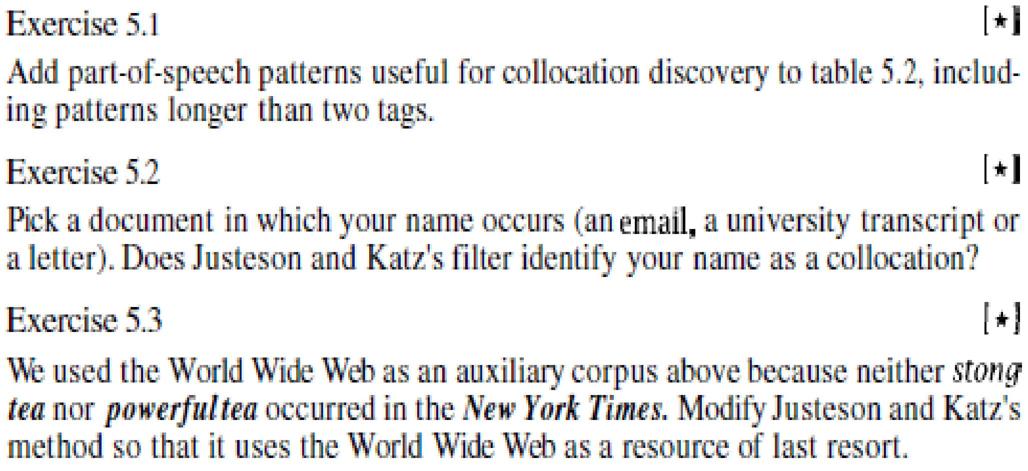
PRACTICAL WORK 8-9. COLLOCATIONS.

FREQUENCY.



Exercise 5.1 [⋆]

Add part-of-speech patterns useful for collocation discovery to table 5.2, including patterns longer than two tags.

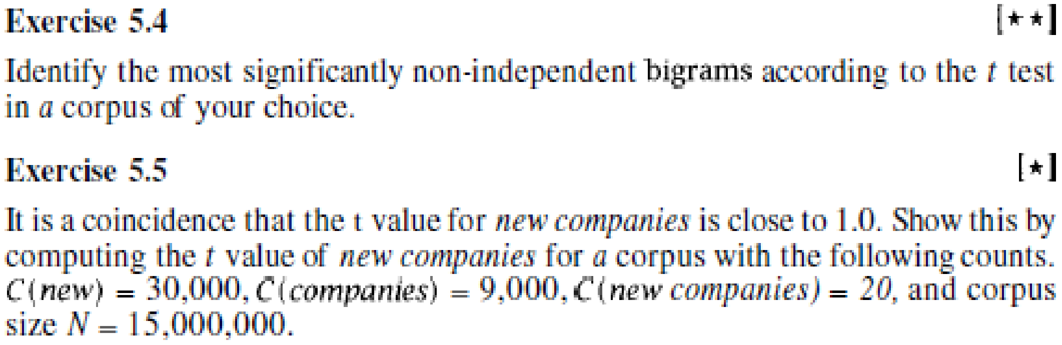
Exercise 5.2 [⋆]

Pick a document in which your name occurs (an email, a university transcript or a letter). Does Justeson and Katz's filter identify your name as a collocation?

Exercise 5.3 [⋆]

We used the World Wide Web as an auxiliary corpus above because neither *stong tea* nor *powerfultea* occurred in the *New York Times.* Modify Justeson and Katz's method so that it uses the World Wide Web as a resource of last resort.

MEAN AND VARIANCE. HYPOTHESIS TESTING.

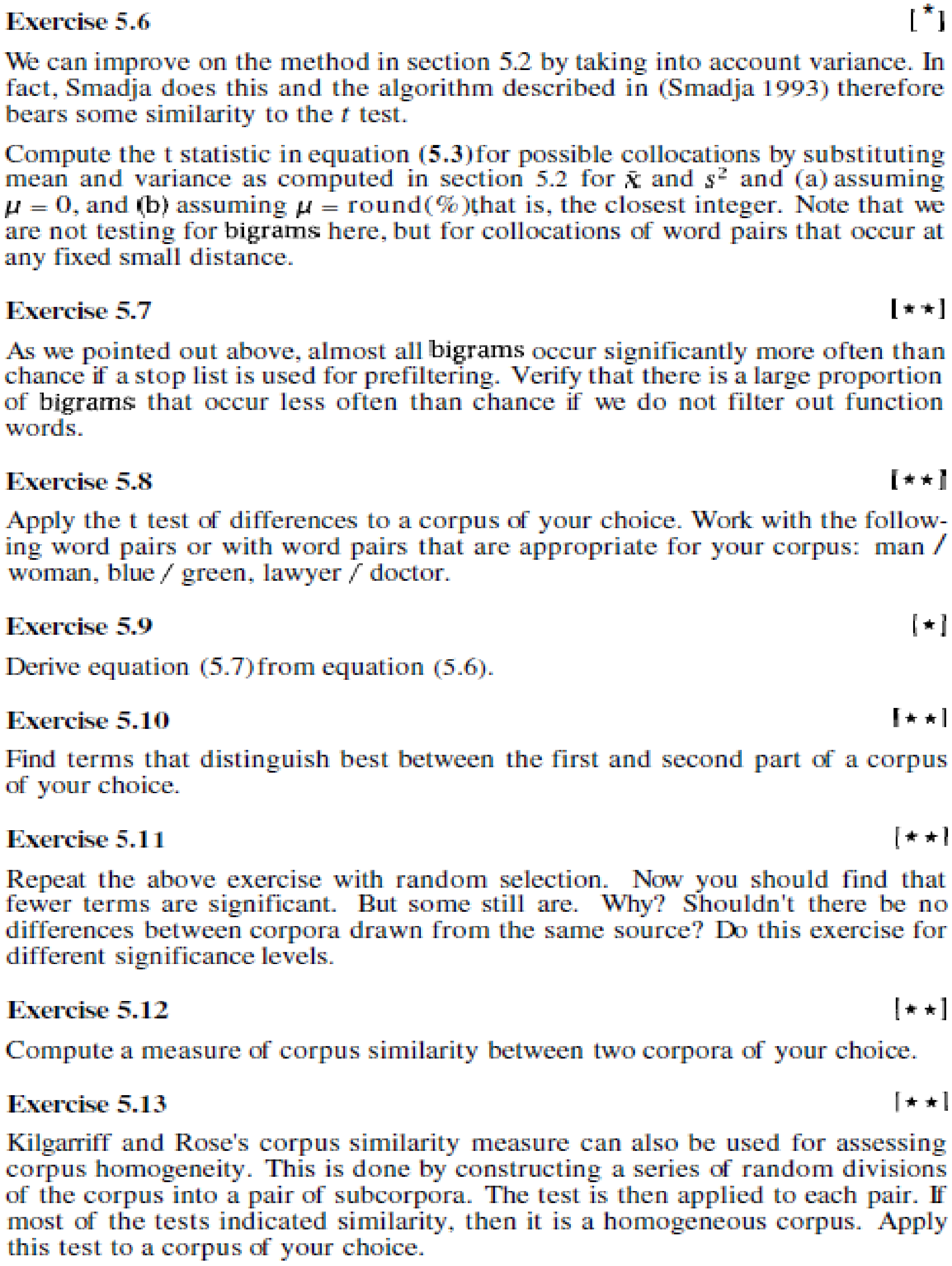


Exercise 5.4 [⋆⋆]

Identify the most significantly non-independent bigrams according to the t test in *a* corpus of your choice.

Exercise 5.5 [⋆]

It is a coincidence that the t value for *new companies* is close to 1.0. Show this by computing the t value of *new companies* for *a* corpus with the following counts. $C(new) = 30{,}000$, $C(companies) = 9{,}000$, $C(new\ companies) = 20$, and corpus size $N = 15{,}000{,}000$.



Exercise 5.6 [*]

We can improve on the method in section 5.2 by taking into account variance. In fact, Smadja does this and the algorithm described in (Smadja 1993) therefore bears some similarity to the t test.

Compute the t statistic in equation (**5.3**) for possible collocations by substituting mean and variance as computed in section 5.2 for $\bar{x}$ and s^2 and (a) assuming $\mu = 0$, and (b) assuming $\mu = \mathrm{round}(\%)$that is, the closest integer. Note that we are not testing for bigrams here, but for collocations of word pairs that occur at any fixed small distance.

Exercise 5.7 [⋆⋆]

As we pointed out above, almost all bigrams occur significantly more often than chance if a stop list is used for prefiltering. Verify that there is a large proportion of bigrams that occur less often than chance if we do not filter out function words.

Exercise 5.8 [⋆⋆]

Apply the t test of differences to a corpus of your choice. Work with the following word pairs or with word pairs that are appropriate for your corpus: man / woman, blue / green, lawyer / doctor.

Exercise 5.9 [⋆]

Derive equation (5.7) from equation (5.6).

Exercise 5.10 [⋆⋆]

Find terms that distinguish best between the first and second part of a corpus of your choice.

Exercise 5.11 [⋆⋆]

Repeat the above exercise with random selection. Now you should find that fewer terms are significant. But some still are. Why? Shouldn't there be no differences between corpora drawn from the same source? Do this exercise for different significance levels.

Exercise 5.12 [⋆⋆]

Compute a measure of corpus similarity between two corpora of your choice.

Exercise 5.13 [⋆⋆]

Kilgarriff and Rose's corpus similarity measure can also be used for assessing corpus homogeneity. This is done by constructing a series of random divisions of the corpus into a pair of subcorpora. The test is then applied to each pair. If most of the tests indicated similarity, then it is a homogeneous corpus. Apply this test to a corpus of your choice.

MUTUAL INFORMATION.

Exercise 5.14 [★★]

Justeson and Katz's part-of-speech filter in section 5.1 can be applied to any of the other methods of collocation discovery in this chapter. Pick one and modify it to incorporate a part-of-speech filter. What advantages does the modified method have?

Exercise 5.15 [★★★]

Design and implement a collocation discovery tool for a translator's workbench. Pick either one method or a combination of methods that the translator can choose from.

Exercise 5.16 [★★★]

Design and implement a collocation discovery tool for a lexicographer's workbench. Pick either one method or a combination of methods that the lexicographer can choose from.

Exercise 5.17 [★★★]

Many news services tag references to companies in their news stories. For example, all references to the *General Electric Company* would be tagged with the same tag regardless of which variant of the name is used (e.g., *GE, General Electric,* or *General Electric Company).* Design and implement a collocation discovery tool for finding company names. How could one partially automate the process of identifying variants?