cur extremely infrequently: over 90% of the word types occur 10 times or less. Nevertheless, very rare words make up a considerable proportion of the text: 12% of the text is words that occur 3 times or less.

Such simple text counts as these can have a use in applications such as cryptography, or to give some sort of indication of style or authorship. But such primitive statistics on the distribution of words in a text are hardly terribly linguistically significant. So towards the end of the chapter we will begin to explore a research avenue that has slightly more linguistic interest. But these primitive text statistics already tell us the reason that Statistical NLP is difficult: it is hard to predict much about the behavior of words that you never or barely ever observed in your corpus. One might initially think that these problems would just go away when one uses a larger corpus, but this hope is not borne out: rather, lots of words that we do not see at all in *Tom* Sawyer will occur – once or twice – in a large corpus. The existence of this long tail of rare words is the basis for the most celebrated early result in corpus linguistics, Zipf's law, which we will discuss next.

1.4.3 Zipf's laws

In his book Human *Behavior and the Principle of Least Effort,* Zipf argues that he has found a unifying principle, the Principle of Least Effort, which underlies essentially the entire human condition (the book even includes some questionable remarks on human sexuality!). The Principle of Least Effort argues that people will act so as to minimize their probable average rate of work (i.e., not only to minimize the work that they would have to do immediately, but taking due consideration of future work that might result from doing work poorly in the short term). The evidence for this theory is certain empirical laws that Zipf uncovered, and his presentation of these laws begins where his own research began, in uncovering certain statistical distributions in language. We will not comment on his general theory here, but will mention some of his empirical language laws.

The famous law: Zipf's law

If we count up how often each word (type) of a language occurs in a large corpus, and then list the words in order of their frequency of occurrence, we can explore the relationship between the frequency of a word f and its position in the list, known as its *rank* r. Zipf's law says that:

RANK

Word	Freq. (f)	Rank (r)	$f \cdot r$	Word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332	turned	51	200	10200
and	2972	2	5944	you'll	30	300	9000
a	1775	3	5235	name	21	400	8400
he	877	10	8770	comes	16	500	8000
but	410	20	8400	group	13	600	7800
be	294	30	8820	lead	11	700	7700
there	222	40	8880	friends	10	800	8000
one	172	50	8600	begin	9	900	8100
about	158	60	9480	family	8	1000	8000
more	138	70	9660	brushed	4	2000	8000
never	124	80	9920	sins	2	3000	6000
Oh	116	90	10440	Could	2	4000	8000
two	104	100	10400	Applausive	1	8000	8000

Table *1.3* Empirical evaluation of Zipf's law on Tom *Sawyer.*

(1.14) $$f \propto \frac{1}{r}$$

or, in other words:

(1.15) There is a constant k such that $f \cdot r = k$

For example, this says that the 50^{th} most common word should occur with three times the frequency of the 150^{th} most common word. This relationship between frequency and rank appears first to have been noticed by Estoup (1916), but was widely publicized by Zipf and continues to bear his name. We will regard this result not actually as a law, but as a roughly accurate characterization of certain empirical facts.

Table 1.3 shows an empirical evaluation of Zipf's law on the basis of Tom Sawyer. Here, Zipf's law is shown to approximately hold, but we note that it is quite a bit off for the three highest frequency words, and further that the product $f \cdot r$ tends to bulge a little for words of rank around 100, a slight bulge which can also be noted in many of Zipf's own studies. Nevertheless, Zipf's law is useful as a rough description of the frequency distribution of words in human languages: there are a few very common words, a middling number of medium frequency words, and many low frequency words. Zipf saw in this a deep significance.

According to his theory both the speaker and the hearer are trying to minimize their effort. The speaker's effort is conserved by having a small vocabulary of common words and the hearer's effort is lessened by having a large vocabulary of individually rarer words (so that messages are less ambiguous). The maximally economical compromise between these competing needs is argued to be the kind of reciprocal relationship between frequency and rank that appears in the data supporting Zipf's law. However, for us, the main upshot of Zipf's law is the practical problem that for most words our data about their use will be exceedingly sparse. Only for a few words will we have lots of examples.

The validity and possibilities for the derivation of Zipf's law is studied extensively by Mandelbrot (1954). While studies of larger corpora sometimes show a closer match to Zipf's predictions than our examples here, Mandelbrot (1954: 12) also notes that "bien que la formule de Zipf donne l'allure générale des courbes, elle en represente très mal les details [although Zipf's formula gives the general shape of the curves, it is very bad in reflecting the details]." Figure 1.1 shows a rank-frequency plot of the words in one corpus (the Brown corpus) on doubly logarithmic axes. Zipf's law predicts that this graph should be a straight line with slope -1. Mandelbrot noted that the line is often a bad fit, especially for low and high ranks. In our example, the line is too low for most low ranks and too high for ranks greater than 10,000.

To achieve a closer fit to the empirical distribution of words, Mandelbrot derives the following more general relationship between rank and frequency:

$$f = P(r + \rho)^{-B} \quad \text{or} \quad \log f = \log P - B \log(r + \rho) \tag{1.16}$$

Here P, B and ρ are parameters of a text, that collectively measure the richness of the text's use of words. There is still a hyperbolic distribution between rank and frequency, as in the original equation (1.14). If this formula is graphed on doubly logarithmic axes, then for large values of r, it closely approximates a straight line descending with slope $-B$, just as Zipf's law. However, by appropriate setting of the other parameters, one can model a curve where the predicted frequency of the most frequent words is lower, while thereafter there is a bulge in the curve: just as we saw in the case of *Tom* Sawyer. The graph in figure 1.2 shows that Mandelbrot's formula is indeed a better fit than Zipf's law for our corpus. The slight bulge in the upper left corner and the larger slope

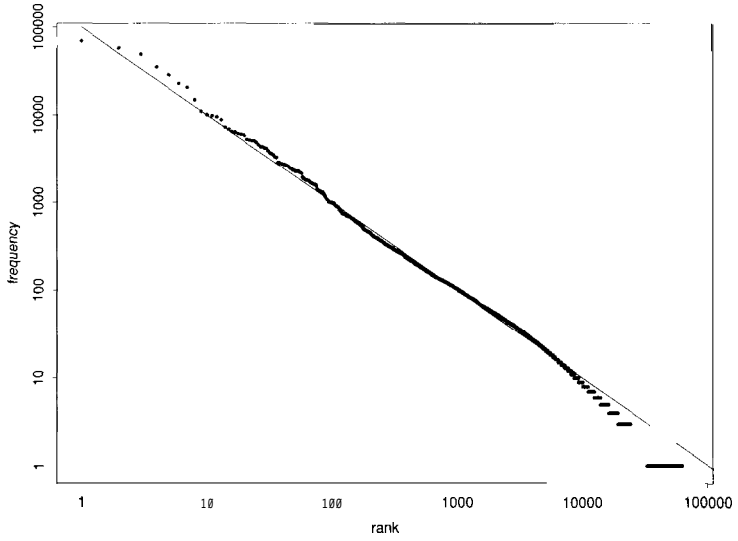

Figure 1.1 Zipf's law. The graph shows rank on the X-axis versus frequency on the Y-axis, using logarithmic scales. The points correspond to the ranks and frequencies of the words in one corpus (the Brown corpus). The line is the relationship between rank and frequency predicted by Zipf for k = 100,000, that is f x r = 100,000.

of $B = 1.15$ model the lowest and highest ranks better than the line in figure 1.1 predicted by Zipf.

If we take $B = 1$ and $\rho = 0$ then Mandelbrot's formula simplifies to the one given by Zipf (see exercise 1.3). Based on data similar to the corpora we just looked at, Mandelbrot argues that Zipf's simpler formula just is not true in general: "lorsque Zipf essayait de représenter tout par cette loi, il essayait d'habiller tout le monde avec des vêtements d'une seule taille [when Zipf tried to represent everything by this (i.e., his) law, he tried to dress everyone with clothes of a single cut]". Nevertheless, Mandelbrot sees the importance of Zipf's work as stressing that there are often phenomena in the world that are not suitably modeled by Gaussian (normal) distributions, that is, 'bell curves,' but by hyperbolic distributions - a fact discovered earlier in the domain of economics by Pareto.

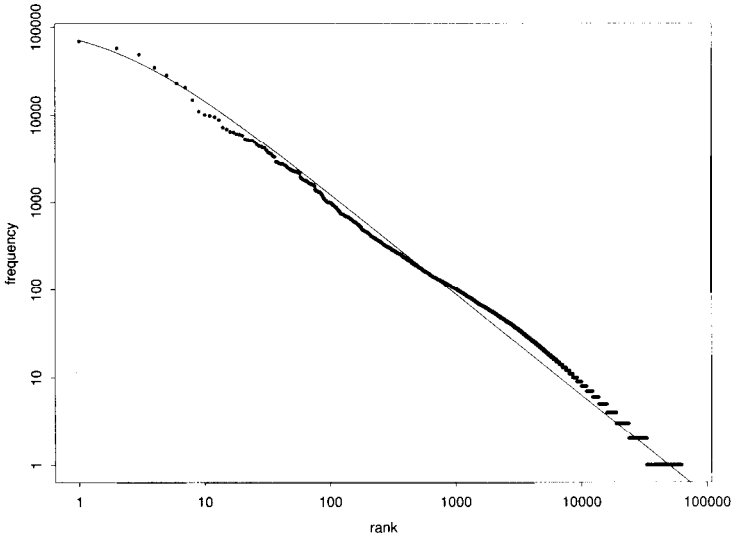

Figure 1.2 Mandelbrot's formula. The graph shows rank on the X-axis versus frequency on the Y-axis, using logarithmic scales. The points correspond to the ranks and frequencies of the words in one corpus (the Brown corpus). The line is the relationship between rank and frequency predicted by Mandelbrot's formula for $P = 10^{5.4}, B = 1.15, \rho = 100$.

Other laws

References to Zipf's law in the Statistical NLP literature invariably refer to the above law, but Zipf actually proposed a number of other empirical laws relating to language which were also taken to illustrate the Principle of Least Effort. At least two others are of some interest to the concerns of Statistical NLP. One is the suggestion that the number of meanings of a word is correlated with its frequency. Again, Zipf argues that conservation of speaker effort would prefer there to be only one word with all meanings while conservation of hearer effort would prefer each meaning to be expressed by a different word. Assuming that these forces are equally strong, Zipf argues that the number of meanings m of a word obeys the law:

(1.17) $m \propto \sqrt{f}$

or, given the previous law, that:

$$m \propto \frac{1}{\sqrt{r}} \quad (1.18)$$

Zipf finds empirical support for this result (in his study, words of frequency rank about 10,000 average about 2.1 meanings, words of rank about 5000 average about 3 meanings, and words of rank about 2000 average about 4.6 meanings).

A second result concerns the tendency of content words to clump. For a word one can measure the number of lines or pages between each occurrence of the word in a text, and then calculate the frequency F of different interval sizes I. For words of frequency at most 24 in a 260,000 word corpus, Zipf found that the number of intervals of a certain size was inversely related to the interval size ($F \propto I^{-p}$, where p varied between about 1 and 1.3 in Zipf's studies). In other words, most of the time content words occur near another occurrence of the same word.

▼ The topic of word senses is discussed in chapter 7, while the clumping of content words is discussed in section 15.3.

Other laws of Zipf's include that there is an inverse relationship between the frequency of words and their length, that the greater the frequency of a word or morpheme, the greater the number of different permutations (roughly, compounds and morphologically complex forms) it will be used in, and yet further laws covering historical change and the frequency of phonemes.

The significance of power laws

As a final remark on Zipf's law, we note that there is a debate on how surprising and interesting Zipf's law and 'power laws' in general are as a description of natural phenomena. It has been argued that randomly generated text exhibits Zipf's law (Li 1992). To show this, we construct a generator that randomly produces characters from the 26 letters of the alphabet and the blank (that is, each of these 27 symbols has an equal chance of being generated next). Simplifying slightly, the probability of a word of length n being generated is $(\frac{26}{27})^n \frac{1}{27}$: the probability of generating a non-blank character n times and the blank after that. One can show that the words generated by such a generator obey a power law of the form Mandelbrot suggested. The key insights are (i) that there are 26 times more words of length $n + 1$ than length n, and (ii) that there is a

constant ratio by which words of length n are more frequent than words of length n + 1. These two opposing trends combine into the regularity of Mandelbrot's law. See exercise 1.4.

There is in fact a broad class of probability distributions that obey power laws when the same procedure is applied to them that is used to compute the Zipf distribution: first counting events, then ranking them according to their frequency (Günter et al. 1996). Seen from this angle, Zipf's law seems less valuable as a characterization of language. But the basic insight remains: what makes frequency-based approaches to language hard is that almost all words are rare. Zipf's law is a good way to encapsulate this insight.

1.4.4 Collocations

COLLOCATION

Lexicographers and linguists (although rarely those of a generative bent) have long been interested in collocations. A collocation is any turn of phrase or accepted usage where somehow the whole is perceived to have an existence beyond the sum of the parts. Collocations include compounds ***(disk drive),*** phrasal verbs ***(make up),*** and other stock phrases (bacon and eggs). They often have a specialized meaning or are idiomatic, but they need not be. For example, at the time of writing, a favorite expression of bureaucrats in Australia is ***international best practice. Now*** there appears to be nothing idiomatic about this expression; it is simply two adjectives modifying a noun in a productive and semantically compositional way. But, nevertheless, the frequent use of this phrase as a fixed expression accompanied by certain connotations justifies regarding it as a collocation. Indeed, any expression that people repeat because they have heard others using it is a candidate for a collocation.

▼ Collocations are discussed in detail in chapter 5. We see later on that collocations are important in areas of Statistical NLP such as machine translation (chapter 13) and information retrieval (chapter 15). In machine translation, a word may be translated differently according to the collocation it occurs in. An information retrieval system may want to index only 'interesting' phrases, that is, those that are collocations.

Lexicographers are also interested in collocations both because they show frequent ways in which a word is used, and because they are multiword units which have an independent existence and probably should appear in a dictionary. They also have theoretical interest: to the extent that most of language use is people reusing phrases and constructions

Frequency	Word 1	Word 2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Table *1.4* Commonest bigram collocations in the New *York Times.*

that they have heard, this serves to de-emphasize the Chomskyan focus on the creativity of language use, and to give more strength to something like a Hallidayan approach that considers language to be inseparable from its pragmatic and social context.

Now collocations may be several words long (such as *international best practice)* or they may be discontinuous (such as make *[something] up),* but let us restrict ourselves to the simplest case and wonder how we can automatically identify contiguous two word collocations. It was mentioned above that collocations tend to be frequent usages. So the first idea to try might be simply to find the most common two word sequences in a text. That is fairly easily done, and, for a corpus of text from the New York *Times (see* page 153), the results are shown in table 1.4. Unfortunately, this method does not seem to succeed very well at capturing the collocations present in the text. It is not surprising that these pairs of words

BIGRAMS (normally referred to as *bigrams*) occur commonly. They simply represent common syntactic constructions involving individually extremely common words. One problem is that we are not normalizing for the frequency of the words that make up the collocation. Given that the, *of*, and *in* are extremely common words, and that the syntax of prepositional and noun phrases means that a determiner commonly follows a preposition, we should expect to commonly see of the and *in the*. But that does not make these word sequences collocations. An obvious next step is to somehow take into account the frequency of each of the words. We will look at methods that do this in chapter 5.

A modification that might be less obvious, but which is very effective, is to *filter* the collocations and remove those that have parts of speech (or syntactic categories) that are rarely associated with interesting collocations. There simply are no interesting collocations that have a preposition as the first word and an article as the second word. The two most frequent patterns for two word collocations are "adjective noun" and "noun noun" (the latter are called noun-noun compounds). Table 1.5 shows which bigrams are selected from the corpus if we only keep adjective-noun and noun-noun bigrams. Almost all of them seem to be phrases that we would want to list in a dictionary – with some exceptions like *last year* and *next year*.

Our excursion into 'collocation discovery' illustrates the back and forth in Statistical NLP between modeling and data analysis. Our initial model was that a collocation is simply a frequent bigram. We analyzed the results we got based on this model, identified problems and then came up with a refined model (collocation = frequent bigram with a particular part-of-speech pattern). This model needs further refinement because of bigrams like *next year* that are selected incorrectly. Still, we will leave our investigation of collocations for now, and continue it in chapter 5.

1.4.5 Concordances

As a final illustration of data exploration, suppose we are interested in the syntactic frames in which verbs appear. People have researched how to get a computer to find these frames automatically, but we can also just use the computer as a tool to find appropriate data. For such purposes, KEY WORD IN CONTEXT people often use a *Key Word In Context* (KWIC) concordancing program which produces displays of data such as the one in figure 1.3. In such a display, all occurrences of the word of interest are lined up beneath

Frequency	Word 1	Word 2	Part-of-speech pattern
11487	New	York	AN
7261	United	States	AN
5412	Los	Angeles	NN
3301	last	year	AN
3191	Saudi	Arabia	NN
2699	last	week	AN
2514	vice	president	AN
2378	Persian	Gulf	AN
2161	San	Francisco	NN
2106	President	Bush	NN
2001	Middle	East	AN
1942	Saddam	Hussein	NN
1867	Soviet	Union	AN
1850	White	House	AN
1633	United	Nations	AN
1337	York	City	NN
1328	oil	prices	NN
1210	next	year	AN
1074	chief	executive	AN
1073	real	estate	AN

Table 1.5 Frequent bigrams after filtering. The most frequent bigrams in the New *York Times* after applying a part-of-speech filter.

```
1   could find a target. The librarian   "showed   off" - running hither and thither w
2  elights in. The young lady teachers   "showed   off" - bending sweetly over pupils
3  ingly. The young gentlemen teachers   "showed   off" with small scoldings and other
4  seeming vexation). The little girls   "showed   off" in various ways, and the littl
5  n various ways, and the little boys   "showed   off" with such diligence that the a
6  t genuwyne?" Tom lifted his lip and    showed   the vacancy. "Well, all right," sai
   is little finger for a pen. Then he    showed   Huckleberry how to make an H and an
8  ow's face was haggard, and his eyes    showed   the fear that was upon him. When he
9  not overlook the fact that Tom even    showed   a marked aversion to these inquests
10 own. Two or three glimmering lights    showed   where it lay, peacefully sleeping,
11 ird flash turned night into day and     showed   every little grass-blade, separate
12  that grew about their feet. And it    showed   three white, startled faces, too. A
13 he first thing his aunt said to him     showed   him that he had brought his sorrows
14 p from her lethargy of distress and     showed   good interest in the proceedings. S
15 ent a new burst of grief from Becky     showed   Tom that the thing in his mind had
16  shudder quiver all through him. He    showed   Huck the fragment of candle-wick pe
```

Figure 1.3 Key Word In Context (KWIC) display for the word *showed*.

NP_{agent} showed off ($PP[with/in]_{manner}$)

NP_{agent} showed ($NP_{recipient}$) ({ $NP_{content}$ | $CP[that]_{content}$ | $VP[inf]_{content}$ | *how* $VP[inf]_{content}$ | $CP[where]_{content}$ } |)

NP_{agent} showed NP[*interest]* $PP[in]_{content}$

NP_{agent} showed NP[*aversion*] $PP[to]_{content}$

Figure 1.4 Syntactic frames for ***showed*** in ***Tom Sawyer.***

one another, with surrounding context shown on both sides. Commonly, KWIC programs allow you to sort the matches by left or right context. However, if we are interested in syntactic frames, rather than particular words, such sorting is of limited use. The data shows occurrences of the word showed within the novel Torn *Sawyer.* There are 5 uses of ***showed off*** (actually all within one paragraph of the text), each in double quotes, perhaps because it was a neologism at the time, or perhaps because Twain considered the expression slang. All of these uses are intransitive, although some take prepositional phrase modifiers. Beyond these, there are four straightforward transitive verb uses with just a direct object (6, 8, 11, 12) – although there are interesting differences between them with 8 being nonagentive, and 12 illustrating a sense of 'cause to be visible.' There is one ditransitive use which adds the person being shown (16). Three examples make who was shown the object NP and express the content either as a that-clause (13, 15) or as a non-finite question-form complement clause (7). One other example has a finite question-form complement clause (10) but omits mention of the person who is shown. Finally two examples have an NP object followed by a prepositional phrase and are quite idiomatic constructions (9, 14): show an ***aversion*** *PP[to]* and show ***an interest*** *PP[in].* But note that while quite idiomatic, they are not completely frozen forms, since in both cases the object noun is productively modified to make a more complex NP. We could systematize the patterns we have found as in figure 1.4.

Collecting information like this about patterns of occurrence of verbs can be useful not only for purposes such as dictionaries for learners of foreign languages, but for use in guiding statistical parsers. A substantial part of the work in Statistical NLP consists (or should consist!) of poring

over large amounts of data, like concordance lines and lists of candidates for collocations. At the outset of a project this is done to understand the important phenomena, later to refine the initial modeling, and finally to evaluate what was achieved.

1.5 Further Reading

Chomsky (1965: 47ff, 1980: 234ff, 1986) discusses the distinction between rationalist and empiricist approaches to language, and presents arguments for the rationalist position. A recent detailed response to these arguments from an 'empiricist' is (Sampson 1997). For people from a generative (computational) linguistics background wondering what Statistical NLP can do for them, and how it relates to their traditional concerns, Abney (1996b) is a good place to start. The observation that there must be a preference for certain kinds of generalizations in order to bootstrap induction was pointed out in the machine learning literature by Mitchell (1980), who termed the preference bias. The work of Firth is highly influential within certain strands of the British corpus linguistics tradition, and is thoroughly covered in (Stubbs 1996). References from within the Statistical NLP community perhaps originate in work from AT&T, see for instance (Church and Mercer 1993: 1). The Hallidayan approach to language is presented in (Halliday 1994).

BIAS

Thorough discussions of *grammaticality* judgements in linguistics are found in (Schütze 1996) and (Cowart 1997). Cowart argues for making use of the judgements of a population of speakers, which is quite compatible with the approach of this book, and rather against the Chomskyan approach of exploring the grammar of a single speaker. A good entry point to the literature on categorical perception is (Harnad 1987).

GRAMMATICALITY

Lauer (199Sb: ch. 3) advocates an approach involving probability distributions over meanings. See the Further Reading of chapter 12 for references to other Statistical NLP work that involves mapping to semantic representations.

The discussion of *kind/sort of* is based on Tabor (1994), which should be consulted for the sources of the citations used. Tabor provides a connectionist model which shows how the syntactic change discussed can be caused by changing frequencies of use. A lot of interesting recent work on gradual syntactic change can be found in the literature on *grammaticalization* (Hopper and Traugott 1993).

GRAMMATICALIZATION

Two proponents of an important role for probabilistic mechanisms in cognition are Anderson (1983, 1990) and Suppes (1984). See (Oaksford and Chater 1998) for a recent collection describing different cognitive architectures, including connectionism. The view that language is best explained as a cognitive phenomenon is the central tenet of cognitive linguistics (Lakoff 1987; Langacker 1987, 19911, but many cognitive linguists would not endorse probability theory as a formalization of cognitive linguistics. See also (Schutze 1997).

The novel Tom *Sawyer* is available in the public domain on the internet, currently from sources including the Virginia Electronic Text Center (see the website).

Zipf's work began with (Zipf 1929), his doctoral thesis. His two major books are (Zipf 1935) and (Zipf 1949). It is interesting to note that Zipf was reviewed harshly by linguists in his day (see, for instance, (Kent 1930) and (Prokosch 1933)). In part these criticisms correctly focussed on the grandiosity of Zipf's claims (Kent (1930: 88) writes: "problems of phonology and morphology are not to be solved en *masse* by one grand general formula"), but they also reflected, even then, a certain ambivalence to the application of statistical methods in linguistics. Nevertheless, prominent American structuralists, such as Martin Joos and Morris Swadesh, did become involved in data collection for statistical studies, with Joos (1936) emphasizing that the question of whether to use statistical methods in linguistics should be evaluated separately from Zipf's particular claims.

As well as (Mandelbrot 1954), Mandelbrot's investigation of Zipf's law is summarized in (Mandelbrot 1983) - see especially chapters 38, 40, and 42. Mandelbrot attributes the direction of his life's work (leading to his well known work on fractals and the Mandelbrot set) to reading a review of (Zipf 1949).

Concordances were first constructed by hand for important literary and religious works. Computer concordancing began in the late 1950s for the purposes of categorizing and indexing article titles and abstracts. Luhn (1960) developed the first computer concordancer and coined the term *KWIC*.

KWIC

1.6 Exercises

Exercise 1.1 [★★ Requires some knowledge of linguistics]

Try to think of some other cases of noncategorical phenomena in language, perhaps related to language change. For starters, look at the following pairs of

sentences, and try to work out the problems they raise. (Could these problems be solved simply by assigning the words to two categories, or is there evidence of mixed categoriality?)

(1.19) a. On the weekend the children had *fun*.

b. That's the *funnest* thing we've done all holidays.

(1.20) a. Do you get much *email* at work?

b. This morning I had *emails* from five clients, all complaining.

Exercise 1.2 [⋆⋆ Probably best attempted after reading chapter 41

Replicate some of the results of section 1.4 on some other piece of text. (Alternatively, you could use the same text that we did so that you can check your work easily. In this case, you should only expect results similar to ours, since the exact numbers depend on various details of what is treated as a word, how case distinctions are treated, etc.)

Exercise 1.3 [⋆]

Show that Mandelbrot's law simplifies to Zipf's law for $B = 1$ and $\rho = 0$.

Exercise 1.4 [⋆⋆]

Construct a table like table 1.3 for the random character generator described above on page 29 (which generates the letters a through z and blank with equal probability of 1/ 27).

Exercise 1.5 [⋆⋆]

Think about ways of identifying collocations that might be better than the methods used in this chapter.

Exercise 1.6 [⋆⋆]

If you succeeded in the above exercise, try the method out and see how well it appears to perform.

Exercise 1.7 [⋆⋆]

Write a program to produce KWIC displays from a text file. Have the user be able to select the word of interest and the size of the surrounding context.