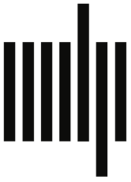
Foundations of Statistical Natural Language Processing

Christopher Manning and Hinrich Schuetze



The MIT Press

From The MIT Press



MITCogNet

© 1999 Massachusetts Institute of Technology
Second printing with corrections 1999
Third printing 2000, fourth printing 2001
Fifth printing with corrections

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

Typeset in 10/13 Lucida Bright by the authors using LaTeX 2ε.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Information

Manning, Christopher D.
Foundations of statistical natural language processing / Christopher D. Manning, Hinrich Schütze.
p. cm.
Includes bibliographical references (p.) and index.
ISBN 0-262-13360-1
1. Computational linguistics—Statistical methods. I. Schütze, Hinrich. II. Title.
P98.5.S83M36 1999
410'.285—dc21 99-21137
CIP

10 9 8 7 6

1 *Introduction*

THE AIM of a linguistic science is to be able to characterize and explain the multitude of linguistic observations circling around us, in conversations, writing, and other media. Part of that has to do with the cognitive side of how humans acquire, produce, and understand language, part of it has to do with understanding the relationship between linguistic utterances and the world, and part of it has to do with understanding the linguistic structures by which language communicates. In order to approach the last problem, people have proposed that there are *rules* which are used to structure linguistic expressions. This basic approach has a long history that extends back at least 2000 years, but in this century the approach became increasingly formal and rigorous as linguists explored detailed grammars that attempted to describe what were well-formed versus ill-formed utterances of a language.

RULES

However, it has become apparent that there is a problem with this conception. Indeed it was noticed early on by Edward Sapir, who summed it up in his famous quote "All grammars leak" (Sapir 1921: 38). It is just not possible to provide an exact and complete characterization of well-formed utterances that cleanly divides them from all other sequences of words, which are regarded as ill-formed utterances. This is because people are always stretching and bending the 'rules' to meet their communicative needs. Nevertheless, it is certainly not the case that the rules are completely ill-founded. Syntactic rules for a language, such as that a basic English noun phrase consists of an optional determiner, some number of adjectives, and then a noun, do capture major patterns within the language. But somehow we need to make things looser, in accounting for the creativity of language use.

This book explores an approach that addresses this problem head on. Rather than starting off by dividing sentences into grammatical and ungrammatical ones, we instead ask, "What are the common patterns that occur in language use?" The major tool which we use to identify these patterns is counting things, otherwise known as statistics, and so the scientific foundation of the book is found in probability theory. Moreover, we are not merely going to approach this issue as a scientific question, but rather we wish to show how statistical models of language are built and successfully used for many natural language processing (NLP) tasks. While practical utility is something different from the validity of a theory, the usefulness of statistical models of language tends to confirm that there is something right about the basic approach.

Adopting a Statistical NLP approach requires mastering a fair number of theoretical tools, but before we delve into a lot of theory, this chapter spends a bit of time attempting to situate the approach to natural language processing that we pursue in this book within a broader context. One should first have some idea about *why* many people are adopting a statistical approach to natural language processing and of *how* one should go about this enterprise. So, in this first chapter, we examine some of the philosophical themes and leading ideas that motivate a statistical approach to linguistics and NLP, and then proceed to get our hands dirty by beginning an exploration of what one can learn by looking at statistics over texts.

1.1 Rationalist and Empiricist Approaches to Language

Some language researchers and many NLP practitioners are perfectly happy to just work on text without thinking much about the relationship between the mental representation of language and its manifestation in written form. Readers sympathetic with this approach may feel like skipping to the practical sections, but even practically-minded people have to confront the issue of what prior knowledge to try to build into their model, even if this prior knowledge might be clearly different from what might be plausibly hypothesized for the brain. This section briefly discusses the philosophical issues that underlie this question.

RATIONALIST

Between about 1960 and 1985, most of linguistics, psychology, artificial intelligence, and natural language processing was completely dominated by a *rationalist* approach. A rationalist approach is characterized

by the belief that a significant part of the knowledge in the human mind is not derived by the senses but is fixed in advance, presumably by genetic inheritance. Within linguistics, this rationalist position has come to dominate the field due to the widespread acceptance of arguments by Noam Chomsky for an innate language faculty. Within artificial intelligence, rationalist beliefs can be seen as supporting the attempt to create intelligent systems by handcoding into them a lot of starting knowledge and reasoning mechanisms, so as to duplicate what the human brain begins with.

POVERTY OF THE STIMULUS

Chomsky argues for this innate structure because of what he perceives as a problem of the *poverty of the stimulus* (e.g., Chomsky 1986: 7). He suggests that it is difficult to see how children can learn something as complex as a natural language from the limited input (of variable quality and interpretability) that they hear during their early years. The rationalist approach attempts to dodge this difficult problem by postulating that the key parts of language are innate – hardwired in the brain at birth as part of the human genetic inheritance.

EMPIRICIST

An *empiricist* approach also begins by postulating some cognitive abilities as present in the brain. The difference between the approaches is therefore not absolute but one of degree. One has to assume some initial structure in the brain which causes it to prefer certain ways of organizing and generalizing from sensory inputs to others, as no learning is possible from a completely blank slate, a *tabula rasa*. But the thrust of empiricist approaches is to assume that the mind does not begin with detailed sets of principles and procedures specific to the various components of language and other cognitive domains (for instance, theories of morphological structure, case marking, and the like). Rather, it is assumed that a baby's brain begins with general operations for association, pattern recognition, and generalization, and that these can be applied to the rich sensory input available to the child to learn the detailed structure of natural language. Empiricism was dominant in most of the fields mentioned above (at least the ones then existing!) between 1920 and 1960, and is now seeing a resurgence. An empiricist approach to NLP suggests that we can learn the complicated and extensive structure of language by specifying an appropriate general language model, and then inducing the values of parameters by applying statistical, pattern recognition, and machine learning methods to a large amount of language use.

Generally in Statistical NLP, people cannot actually work from observing a large amount of language use situated within its context in the

world. So, instead, people simply use texts, and regard the textual context as a surrogate for situating language in a real world context. A body of texts is called a *corpus* – *corpus* is simply Latin for 'body,' and when you have several such collections of texts, you have *corpora*. Adopting such a corpus-based approach, people have pointed to the earlier advocacy of empiricist ideas by the British linguist J. R. Firth, who coined the slogan "You shall know a word by the company it keeps" (Firth 1957: 11). However an empiricist corpus-based approach is perhaps even more clearly seen in the work of American structuralists (the 'post-Bloomfieldians'), particularly Zellig Harris. For example, (Harris 1951) is an attempt to find discovery procedures by which a language's structure can be discovered automatically. While this work had no thoughts to computer implementation, and is perhaps somewhat computationally naive, we find here also the idea that a good grammatical description is one that provides a compact representation of a corpus of texts.

CORPUS
CORPORA
AMERICAN STRUCTURALISTS

It is not appropriate to provide a detailed philosophical treatment of scientific approaches to language here, but let us note a few more differences between rationalist and empiricist approaches. Rationalists and empiricists are attempting to describe different things. Chomskyan (or *generative*) linguistics seeks to describe the language module of the human mind (the I-language) for which data such as texts (the E-language) provide only indirect evidence, which can be supplemented by native speaker intuitions. Empiricist approaches are interested in describing the E-language as it actually occurs. Chomsky (1965: 3–4) thus makes a crucial distinction between *linguistic competence*, which reflects the knowledge of language structure that is assumed to be in the mind of a native speaker, and *linguistic performance* in the world, which is affected by all sorts of things such as memory limitations and distracting noises in the environment. Generative linguistics has argued that one can isolate linguistic competence and describe it in isolation, while empiricist approaches generally reject this notion and want to describe actual use of language.

GENERATIVE LINGUISTICS
LINGUISTIC COMPETENCE
LINGUISTIC PERFORMANCE

This difference underlies much of the recent revival of interest in empiricist techniques for computational work. During the second phase of work in artificial intelligence (roughly 1970–1989, say) people were concerned with the science of the mind, and the best way to address that was seen as building small systems that attempted to behave intelligently. This approach identified many key problems and approaches that are

still with us today, but the work can be criticized on the grounds that it dealt only with very small (often pejoratively called 'toy') problems, and often did not provide any sort of objective evaluation of the general efficacy of the methods employed. Recently, people have placed greater emphasis on engineering practical solutions. Principally, they seek methods that can work on raw text as it exists in the real world, and objective comparative evaluations of how well different methods work. This new emphasis is sometimes reflected in naming the field 'Language Technology' or 'Language Engineering' instead of NLP. As we will discuss below, such goals have tended to favor Statistical NLP approaches, because they are better at automatic learning (*knowledge induction*), better at disambiguation, and also have a role in the science of linguistics.

INDUCTION

Finally, Chomskyan linguistics, while recognizing certain notions of competition between principles, depends on *categorical* principles, which sentences either do or do not satisfy. In general, the same was true of American structuralism. But the approach we will pursue in Statistical NLP draws from the work of Shannon, where the aim is to assign probabilities to linguistic events, so that we can say which sentences are 'usual' and 'unusual'. An upshot of this is that while Chomskyan linguists tend to concentrate on categorical judgements about very rare types of sentences, Statistical NLP practitioners are interested in good descriptions of the associations and preferences that occur in the totality of language use. Indeed, they often find that one can get good real world performance by concentrating on common types of sentences.

CATEGORICAL

1.2 Scientific Content

Many of the applications of the methods that we present in this book have a quite *applied* character. Indeed, much of the recent enthusiasm for statistical methods in natural language processing derives from people seeing the prospect of statistical methods providing practical solutions to real problems that have eluded solution using traditional NLP methods. But if statistical methods were just a practical engineering approach, an approximation to difficult problems of language that science has not yet been able to figure out, then their interest to us would be rather limited. Rather, we would like to emphasize right at the beginning that there are clear and compelling scientific reasons to be interested in the frequency

with which linguistic forms are used, in other words, statistics, as one approaches the study of language.

1.2.1 Questions that linguistics should answer

What questions does the study of language concern itself with? As a start we would like to answer two basic questions:

- What kinds of things do people say?
- What do these things say/ask/request about the world?

From these two basic questions, attention quickly spreads to issues about how knowledge of language is acquired by humans, and how they actually go about generating and understanding sentences in real time. But let us just concentrate on these two basic questions for now. The first covers all aspects of the structure of language, while the second deals with semantics, pragmatics, and discourse – how to connect utterances with the world. The first question is the bread and butter of corpus linguistics, but the patterns of use of a word can act as a surrogate for deep understanding, and hence can let us also address the second question using corpus-based techniques. Nevertheless patterns in corpora more easily reveal the syntactic structure of a language, and so the majority of work in Statistical NLP has dealt with the first question of what kinds of things people say, and so let us begin with it here.

COMPETENCE
GRAMMAR

GRAMMATICALITY

How does traditional (structuralist/generative) linguistics seek to answer this question? It abstracts away from any attempt to describe the kinds of things that people usually say, and instead seeks to describe a *competence grammar* that is said to underlie the language (and which generative approaches assume to be in the speaker's head). The extent to which such theories approach the question of what people say is merely to suggest that there is a set of sentences – grammatical sentences – which are licensed by the competence grammar, and then other strings of words are ungrammatical. This concept of *grammaticality* is meant to be judged purely on whether a sentence is structurally well-formed, and not according to whether it is the kind of thing that people would say or whether it is semantically anomalous. Chomsky gave *Colorless green ideas sleep furiously* as an example of a sentence that is grammatical, al-

though semantically strange and not the kind of thing you would expect people to say. Syntactic grammaticality is a categorical binary choice.[1]

Now, initially, a distinction between grammatical and ungrammatical sentences does not seem so bad. We immediately notice when a non-native speaker says something really wrong – something ungrammatical – and we are able to correct such sentences to grammatical ones. In contrast, except when there are bad speech errors, a native speaker normally produces grammatical sentences. But there are at least two reasons why we should seek more. Firstly, while maintaining a binary split between grammatical and ungrammatical sentences may seem plausible in simple cases, it becomes increasingly far-fetched as we extend our investigation. Secondly, regardless of this, there are many reasons to be interested in the frequency with which different sentences and sentence types are used, and simply dividing sentences into grammatical and ungrammatical sentences gives no information about this. For instance, very often non-native speakers say or write things that are not in any way syntactically ungrammatical, but just somehow subtly odd. Here's an example from a student essay:

(1.1) In addition to this, she insisted that women were regarded as a different existence from men unfairly.

We might respond to this passage by saying that we can understand the message, but it would sound better expressed slightly differently. This is a statement about the *conventionality* of certain modes of expression. But a convention is simply a way in which people frequently express or do something, even though other ways are in principle possible.

CONVENTIONALITY

The fact that sentences do not divide neatly into two sets – grammatical and ungrammatical ones – is well known to anyone who has been in linguistics for a while. For many of the complicated sentences of interest to theoretical linguistics, it is difficult for human beings to decide whether they are grammatical or not. For example, try your hand at judging the grammaticality of the following sentences drawn (not at random)

1. Some versions of Chomsky's 1980s theory, Government-Binding theory (GB), provide a minor degree of gradedness by suggesting that sentences that disobey some constraints are only sort of weird while ones that disobey other constraints are truly horrible, but the formal theory, in GB and elsewhere, provides little support for these notions. Linguists generally rely on an informal system of stars and question marks for initially grading sentences (where * (ungrammatical) > ?* > ?? > ? (questionable)), but these gradations are converted into a binary grammatical/ungrammatical distinction when people try to develop the principles of grammar.

from (van Riemsdijk and Williams 1986) – a textbook, not even a research paper – before peeking at the answers in the footnote.[2]

(1.2) a. John I believe Sally said Bill believed Sue saw.

b. What did Sally whisper that she had secretly read?

c. John wants very much for himself to win.

d. (Those are) the books you should read before it becomes difficult to talk about.

e. (Those are) the books you should read before talking about becomes difficult.

f. Who did Jo think said John saw him?

g. That a serious discussion could arise here of this topic was quite unexpected.

h. The boys read Mary's stories about each other.

We find that most people disagree with more than one of van Riemsdijk and Williams's claims about which sentences are grammatical. This result raises real questions about what, if anything, generative linguistics is describing.

This difficulty has led to many statements in the linguistics literature about judgements being difficult, or the facts quite obscure, as if somehow there is a categorical answer to whether each sentence is grammatical, but it is hard for human beings to work out what that answer is. Yet, despite these manifest difficulties, most of theoretical linguistics continues to work in a framework that defines such observations to be out of the realm of interest (relegating them to performance effects). We believe that this is unsustainable. On the other hand, it must be noticed that most simple sentences are either clearly acceptable or unacceptable and we would want our theory to be able to account for this observation. Perhaps the right approach is to notice the parallel with other cases of *categorical perception* that have been described in the psychological literature. For instance, although the timing of voicing onset which differentiates a /p/ sound from a /b/ sound is a continuous variable (and its typical

CATEGORICAL PERCEPTION

2. Answers: a. OK, b. bad, c. OK, d. OK, e. bad, f. OK, g. OK, h. bad.

value varies between languages), human beings perceive the results categorically, and this is why a theory of phonology based on categorical phonemes is largely viable, despite all the movements and variations in phonological production occurring in a continuous space. Similarly for syntax, a categorical theory may suffice for certain purposes. Nevertheless, we would argue that the difficulties in giving grammaticality judgements to complex and convoluted sentences show the implausibility of extending a binary distinction between grammatical and ungrammatical strings to all areas of language use.

1.2.2 Non-categorical phenomena in language

But beyond the above difficulties in giving grammaticality judgements, if we peek into the corners of language, we see clear evidence of failures of categorical assumptions, and circumstances where considerations of frequency of use are essential to understanding language. This suggests that while a *categorical view of language* may be sufficient for many purposes, we must see it as an approximation that also has its limitations (just as Newtonian physics is good for many purposes but has its limits).[3]

CATEGORICAL VIEW OF LANGUAGE

One source of data on non-categorical phenomena in language is to look at the history of language change (others are looking at sociolinguistic variation and competing hypotheses during language acquisition). Over time, the words and syntax of a language change. Words will change their meaning and their part of speech. For instance, English *while* used to be exclusively a noun meaning 'time,' a usage that survives mainly in a few fixed phrases such as *to take a while*, but changed to be mainly used as a complementizer introducing subordinate clauses (*While you were out, ...*). It doesn't make sense to say that categorically until some day in 1742 *while* was only a noun and then it became a complementizer – even if this claim is only being made for certain speakers rather than the speech community as a whole. Rather, one would expect a gradual change. One hypothesis is that if the frequency of use of a word in various contexts gradually changes so that it departs from the typical profile of use of words in the category to which it formerly belonged, and rather its profile of use comes to more resemble words of another category, then

3. Readers not familiar with linguistics and NLP may have trouble understanding this section and may wish to skip it, but to return to it after reading chapter 3. The historical examples include various archaic spellings – the standardization of English spelling is a relatively modern phenomenon. Reading them aloud is often helpful for decoding them.

it will come to be reanalyzed as a word of that different category. During the period of change, one would expect to see evidence of noncategorical behavior.

Blending of parts of speech: *near*

At first blush it appears that the word *near* can be used either as an adjective as in (1.3a) or as a preposition (1.3b):

(1.3) a. We will review that decision in the near future.

b. He lives near the station.

Evidence for *near* as an adjective includes its position between a determiner and noun as in (1.3a) – a classic adjective position – and the fact that it can form an adverb by adding *-ly*: *We nearly lost.* Evidence for *near* as a preposition includes that it can head the locative phrase complements of verbs like *live* as in (1.3b) – a classic role for prepositions, and that such a phrase can be modified by *right*, which is normally restricted to modifying prepositional phrases: *He lives right near the station* (cf. *He swam right across the lake* vs. *??That's a right red car*). So far, though, this data is not that surprising: many words in English seem to have multiple parts of speech. For example, many words are both nouns and verbs, such as *play*: *They saw a play* vs. *They play lacrosse on Thursdays.* But the interesting thing is that *near* can simultaneously show adjective properties and preposition properties, and thus appears to behave as a category blend. This happens in sentences like:

(1.4) a. He has never been nearer the center of the financial establishment.

b. We live nearer the water than you thought.

Realization in the comparative form (*nearer*) is a hallmark of adjectives (and adverbs). Other categories do not form comparatives and superlatives.[4] On the other hand, grammatical theory tells us that adjectives and nouns do not take direct objects, hence we have to insert prepositions

4. The thoughtful reader might note that some prepositions do have related forms ending in *-er* which are perhaps related to comparatives (*upper, downer, inner, outer*), but we note that none of these prepositions have a superlative that is formed in analogy to regular adjectival superlatives, as *near* does (that is, *nearest*), and that none of these other forms in *-er* can be used in preposition-like uses. We cannot say: **John lives inner Sydney than Fred.*

after adjectives and say *unsure **of** his beliefs* or *convenient **for** people who work long hours*. In this sense *nearer* is behaving like a preposition by heading a locative phrase and taking a direct object. Thus in these sentences *nearer* is simultaneously showing properties of adjectives and prepositions that are not available to the other category. Hence it is exhibiting a blended status somewhere between these two parts of speech, which are normally taken as categorically distinct.

Language change: *kind of* and *sort of*

New uses for the word sequences *kind of* and *sort of* present a convincing example of how different frequencies of use in certain constructions can lead to what is apparently categorical change. In modern English, the expressions *sort of* and *kind of* have at least two distinct uses. In one, *sort* or *kind* functions as a noun with *of* as a following preposition introducing a prepositional phrase, as in sentences such as *What sort of animal made these tracks?* But there is another usage in which these expressions can best be thought of as degree modifiers, akin to *somewhat* or *slightly*:

(1.5) a. We are kind of hungry.

b. He sort of understood what was going on.

We can tell that *kind/sort of* is not behaving as a normal noun preposition sequence here because it is appearing in contexts – such as between the subject noun phrase and the verb – where normally one cannot insert a noun-preposition sequence (for example, one cannot say **He variety of understood what was going on*).

Historically, *kind* and *sort* were clearly nouns. Among other things, they could be preceded by a determiner and followed by a PP:

(1.6) a. A nette sent in to the see, and of alle kind of fishis gedrynge. [1382]

b. I knowe that sorte of men ryght well. [1560]

Unambiguous degree modifier uses did not appear until the nineteenth century:

(1.7) a. I kind of love you, Sal—I vow. [1804]

b. It sort o' stirs one up to hear about old times. [1833]

It does not appear that this new construction was borrowed from another language. Rather it appears to be a language internal development. How could this innovation have come about?

A plausible hypothesis is to notice that when we have *kind/sort of* preceding an adjective, then it is actually ambiguous between these two readings:

(1.8) a. [NP a [kind] [PP of [NPdense rock]]]

b. [NP a [AP [MOD kind of] dense] rock]

And what one finds is that between the sixteenth and the nineteenth century, there was a significant rise in the use of *kind/sort of* in this [Det {*sort/kind*} of AdjP N] frame:

(1.9) a. Their finest and best, is a kind of course red cloth. [c. 1600]

b. But in such questions as the present, a hundred contradictory views may preserve a kind of imperfect analogy. [1743]

(Note that *course* is here a variant spelling of *coarse*.) In this environment, *sort/kind of* fills a slot that could be occupied by a noun head followed by a preposition, but it also fills a slot that could be occupied by a degree modifier (with a different syntactic structure). As this usage became more common, *kind/sort of* was more commonly being used in a typical degree modifier slot; in other words, it grew to look syntactically more like a degree modifier. Moreover, the semantics of these particular nouns was such that they could easily be thought of as degree modifiers. This frequency change seems to have driven a change in syntactic category, and in time the use of *kind/sort of* was extended to other contexts such as modifying verb phrases.

The general point here is that while language change can be sudden (due to either external or internal factors), it is generally gradual. The details of gradual change can only be made sense of by examining frequencies of use and being sensitive to varying strengths of relationships, and this type of modeling requires statistical, as opposed to categorical, observations.

Although there have only been a few attempts to use Statistical NLP for explaining complex linguistic phenomena, what is exciting about the subject matter of this book from the point of view of theoretical linguistics is that this new way of looking at language may be able to account for

things such as non-categorical phenomena and language change much better than anything existing today.

1.2.3 Language and cognition as probabilistic phenomena

A more radical argument for probability as part of a scientific understanding of language is that human cognition is probabilistic and that language must therefore be probabilistic too since it is an integral part of cognition. A frequent response to our previous examples of non-categorical phenomena in language is that they are marginal and rare. Most sentences are either clearly grammatical or clearly ungrammatical. And most of the time, words are used in only one part of speech, without blending. But if language and cognition as a whole are best explained probabilistically, then probability theory must be a central part of an explanatory theory of language.

The argument for a probabilistic approach to cognition is that we live in a world filled with uncertainty and incomplete information. To be able to interact successfully with the world, we need to be able to deal with this type of information. Suppose you want to determine whether it is safe to wade through a river. You see that the water is flowing slowly, so probably it won't drag you away. You are pretty certain that no piranhas or alligators live in this area. You integrate all this information in evaluating how safe it is to cross the river. Now, if someone tells you, "the water is only knee-deep if you walk towards that tall tree over there", then this linguistic information will be just one more source of information to incorporate. Processing the words, forming an idea of the overall meaning of the sentence, and weighing it in making a decision is no different in principle from looking at the current, forming an idea of the speed of the water, and taking this sensory information into account. So the gist of this argument is that the cognitive processes used for language are identical or at least very similar to those used for processing other forms of sensory input and other forms of knowledge. These cognitive processes are best formalized as probabilistic processes or at least by means of some quantitative framework that can handle uncertainty and incomplete information.

The facts of language often look quite different depending on whether or not one is sympathetic to an important role for quantitative methods in linguistics. A famous example is Chomsky's dictum that probability theory is inappropriate for formalizing the notion of *grammaticality*.

GRAMMATICALITY

He argued that computing the probability of sentences from a corpus of utterances would assign the same low probability to all unattested sentences, grammatical and ungrammatical ones alike, and hence not account for linguistic productivity (Chomsky 1957: 16). This argument only makes sense if one has a bias against probabilistic representation of concepts in general. Consider the cognitive representation of the concept *tall*. Suppose you see a man who is seven feet tall and it is the first person you've ever seen of that height. You will easily recognize this person as a *tall* man, not as an uncategorizable man. Similarly, it will be easy for you to recognize a person of another unattested height, say four feet, as definitely not *tall*. In this book, we will look at probabilistic models that can easily learn and represent this type of regularity and make the right judgement for unattested examples. Indeed, a major part of Statistical NLP is deriving good probability estimates for unseen events. The premise that all unattested instances will be treated alike in a probabilistic framework does not hold.

We believe that much of the skepticism towards probabilistic models for language (and for cognition in general) stems from the fact that the well-known early probabilistic models (developed in the 1940s and 1950s) are extremely simplistic. Because these simplistic models clearly do not do justice to the complexity of human language, it is easy to view probabilistic models in general as inadequate. One of the insights we hope to promote in this book is that complex probabilistic models can be as explanatory as complex non-probabilistic models – but with the added advantage that they can explain phenomena that involve the type of uncertainty and incompleteness that is so pervasive in cognition in general and in language in particular.

These issues relate to the treatment of semantics in Statistical NLP. We mentioned earlier that most existing work in Statistical NLP has concentrated on the lower levels of grammatical processing, and people have sometimes expressed skepticism as to whether statistical approaches can ever deal with meaning. But the difficulty in answering this question is mainly in defining what 'meaning' is! It is often useful in practice if 'meaning' is viewed as symbolic expressions in some language, such as when translating English into a database query language like SQL. This sort of translation can certainly be done using a Statistical NLP system (we discuss the process of translation in chapter 13). But from a Statistical NLP perspective, it is more natural to think of meaning as residing in the distribution of contexts over which words and utterances are used.

USE THEORY OF MEANING

Philosophically, this brings us close to the position adopted in the later writings of Wittgenstein (that is, Wittgenstein 1968), where the meaning of a word is defined by the circumstances of its use (a *use theory of meaning*) – see the quotations at the beginning of the chapter. Under this conception, much of Statistical NLP research directly tackles questions of meaning.

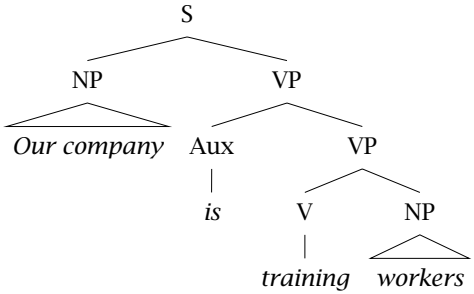
1.3 The Ambiguity of Language: Why NLP Is Difficult

An NLP system needs to determine something of the structure of text – normally at least enough that it can answer "Who did what to whom?" Conventional parsing systems try to answer this question only in terms of possible structures that could be deemed grammatical for some choice of words of a certain category. For example, given a reasonable grammar, a standard NLP system will say that sentence (1.10) has 3 syntactic analyses, often called *parses*:

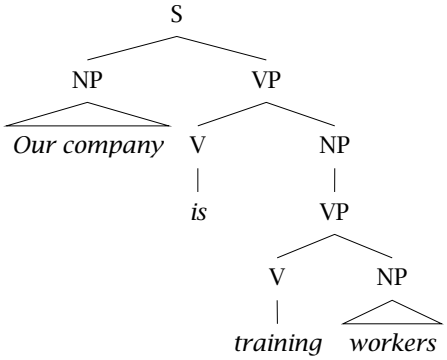
(1.10) Our company is training workers.

The three differing parses might be represented as in (1.11):

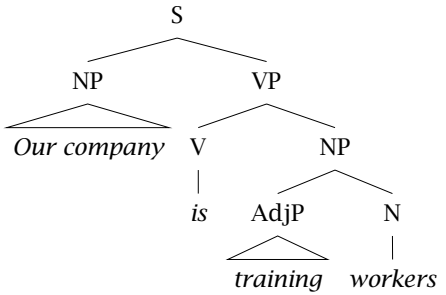
(1.11) a.



b.



c.



There is (a), the one humans perceive, where *is training* is the verb group, and two others with *is* as the main verb: in (b) the rest is a gerund (cf. *Our problem is training workers*), while in (c) *training* modifies *workers* (cf. *Those are training wheels*). The last two parses are semantically anomalous, but in most current systems semantic analysis is done only after syntactic analysis (if at all). This means that, as sentences get longer and grammars get more comprehensive, such ambiguities lead to a terrible multiplication of parses. For instance, Martin et al. (1987) report their system giving 455 parses for the sentence in (1.12):[5]

(1.12) List the sales of the products produced in 1973 with the products produced in 1972.

Therefore, a practical NLP system must be good at making disambiguation decisions of word sense, word category, syntactic structure, and semantic scope. But the goal of maximizing coverage while minimizing resultant ambiguity is fundamentally inconsistent with symbolic NLP systems, where extending the coverage of the grammar to obscure constructions simply increases the number of undesired parses for common sentences and vice versa. Furthermore, experience with AI approaches to parsing and disambiguation, which seek models with deep understanding, has shown that hand-coded syntactic constraints and preference rules are time consuming to build, do not scale up well, and are brittle in the face of the extensive use of metaphor in language (Lakoff 1987).

SELECTIONAL RESTRICTIONS

For instance a traditional approach is to use *selectional restrictions*, and say, for example, that a verb like *swallow* requires an animate being as its subject and a physical object as its object. But such a restriction would disallow common and straightforward metaphorical extensions of the usage of *swallow* such as these:

5. See also Church and Patil (1982) for similar examples.

(1.13) a. I swallowed his story, hook, line, and sinker.

b. The supernova swallowed the planet.

Disambiguation strategies that rely on manual rule creation and hand-tuning produce a knowledge acquisition bottleneck, and still perform poorly when evaluated on naturally occurring text.

A Statistical NLP approach seeks to solve these problems by automatically learning lexical and structural preferences from corpora. Rather than parsing solely using syntactic categories, such as part of speech labels, we recognize that there is a lot of information in the relationships between words, that is, which words tend to group with each other. This collocational knowledge can be exploited as a window onto deeper semantic relationships. In particular, the use of statistical models offers a good solution to the ambiguity problem: statistical models are robust, generalize well, and behave gracefully in the presence of errors and new data. Thus Statistical NLP methods have led the way in providing successful disambiguation in large scale systems using naturally occurring text. Moreover, the parameters of Statistical NLP models can often be estimated automatically from text corpora, and this possibility of automatic learning not only reduces the human effort in producing NLP systems, but raises interesting scientific issues regarding human language acquisition.

1.4 Dirty Hands

1.4.1 Lexical resources

So much for motivation. How does one actually proceed? Well, first of all, one needs to get one's hands on some *lexical resources*: machine-readable text, dictionaries, thesauri, and also tools for processing them. We will briefly introduce a few important ones here since we will be referring to them throughout the book. You can consult the website for more information on how to actually get your hands on them.

LEXICAL RESOURCES

BROWN CORPUS

BALANCED CORPUS

The *Brown corpus* is probably the most widely known corpus. It is a tagged corpus of about a million words that was put together at Brown university in the 1960s and 1970s. It is a *balanced corpus*. That is, an attempt was made to make the corpus a representative sample of American English at the time. Genres covered are press reportage, fiction, scientific text, legal text, and many others. Unfortunately, one has to pay to obtain the Brown corpus, but it is relatively inexpensive for research

purposes. Many institutions with NLP research have a copy available, so ask around. The *Lancaster-Oslo-Bergen* (*LOB*) *corpus* was built as a British English replication of the Brown corpus.

LANCASTER-OSLO-BERGEN CORPUS

SUSANNE CORPUS

The *Susanne corpus* is a 130,000 word subset of the Brown corpus, which has the advantage of being freely available. It is also annotated with information on the syntactic structure of sentences – the Brown corpus only disambiguates on a word-for-word basis. A larger corpus of syntactically annotated (or parsed) sentences is the *Penn Treebank*. The text is from the *Wall Street Journal*. It is more widely used, but not available for free.

PENN TREEBANK

CANADIAN HANSARDS

BILINGUAL CORPUS

PARALLEL TEXTS

The *Canadian Hansards*, the proceedings of the Canadian parliament, are the best known example of a *bilingual corpus*, a corpus that contains *parallel texts* in two or more languages that are translations of each other. Such parallel texts are important for statistical machine translation and other cross-lingual NLP work. The Hansards are another resource that one has to pay for.

WORDNET

SYNSET

In addition to texts, we also need dictionaries. *WordNet* is an electronic dictionary of English. Words are organized into a hierarchy. Each node consists of a *synset* of words with identical (or close to identical) meanings. There are also some other relations between words that are defined, such as meronymy or part-whole relations. WordNet is free and can be downloaded from the internet.

▼ More details on corpora can be found in chapter 4.

1.4.2 Word counts

Once we have downloaded some text, there are a number of quite interesting issues in its low-level representation, classification, and processing. Indeed, so many that chapter 4 is devoted to these questions. But for the moment, let us suppose that our text is being represented as a list of words. For the investigation in this section, we will be using Mark Twain's *Tom Sawyer*.

There are some obvious first questions to ask. What are the most common words in the text? The answer is shown in table 1.1. Notice how this list is dominated by the little words of English which have important grammatical roles, and which are usually referred to as *function words*, such as determiners, prepositions, and complementizers. The one really exceptional word in the list is *Tom* whose frequency clearly reflects the text that we chose. This is an important point. In general the results one

FUNCTION WORDS

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Table 1.1 Common words in *Tom Sawyer*.

gets depends on the corpus or sample used. People use large and varied samples to try to avoid anomalies like this, but in general the goal of using a truly 'representative' sample of all of English usage is something of a chimera, and the corpus will reflect the materials from which it was constructed. For example, if it includes material from linguistics research papers, then words like *ergativity, causativize*, and *lexicalist* may well occur, but otherwise they are unlikely to be in the corpus at all, no matter how large it is.

How many words are there in the text? This question can be interpreted in two ways. The question about the sheer length of the text is distinguished by asking how many *word tokens* there are. There are 71,370. So this is a very small corpus by any standards, just big enough to illustrate a few basic points. Although *Tom Sawyer* is a reasonable length novel, it is somewhat less than half a megabyte of online text, and for broad coverage statistical grammars we will often seek collections of text that are orders of magnitude larger. How many different words, or in other words, how many *word types* appear in the text? There are 8,018. This is actually quite a small number for a text its size, and presumably reflects the fact that *Tom Sawyer* is written in a colloquial style for chil-

WORD TOKENS

WORD TYPES

Word Frequency	Frequency of Frequency
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11–50	540
51–100	99
> 100	102

Table 1.2 Frequency of frequencies of word types in *Tom Sawyer*.

TOKENS
TYPES

dren (for instance, a sample of newswire the same size contained slightly over 11,000 word types). In general in this way one can talk about *tokens*, individual occurrences of something, and *types*, the different things present. One can also calculate the ratio of tokens to types, which is simply the average frequency with which each type is used. For *Tom Sawyer*, it is 8.9.[6]

The above statistics tell us that words in the corpus occur 'on average' about 9 times each. But one of the greatest problems in Statistical NLP is that word types have a very uneven distribution. Table 1.2 shows how many word types occur with a certain frequency. Some words are very common, occurring over 700 times and therefore individually accounting for over 1% of the words in the novel (there are 12 such words in table 1.1). Overall, the most common 100 words account for slightly over half (50.9%) of the word tokens in the text. On the other extreme, note that almost half (49.8%) of the word types occur only once in the corpus.

HAPAX LEGOMENA

Such words are referred to as *hapax legomena*, Greek for 'read only once.' Even beyond these words, note that the vast majority of word types oc-

6. This ratio is not a valid measure of something like 'text complexity' just by itself, since the value varies with the size of the text. For a valid comparison, one needs to normalize the lengths of the texts, such as by calculating the measure over windows of 1,000 words.

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 représente très mal les détails [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

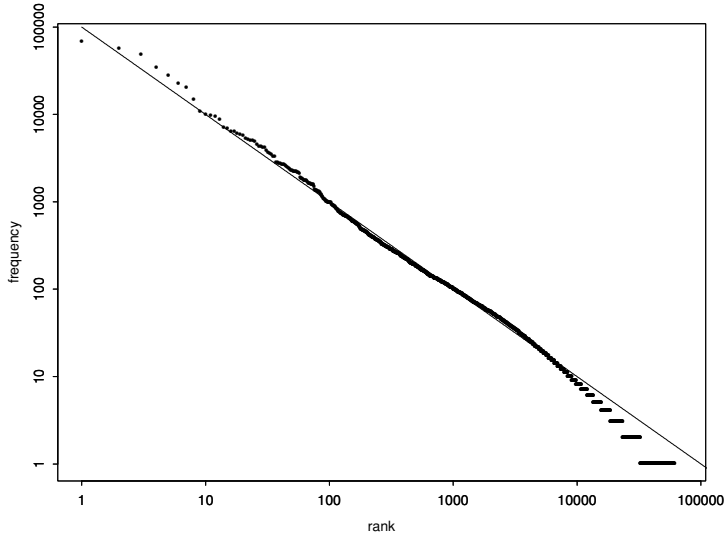


Figure 1.1 Zipf's law. The graph shows rank on the X-axis versus frequency on the Y-axis, using logarithmic scales. The points correspond to the ranks and frequencies of the words in one corpus (the Brown corpus). The line is the relationship between rank and frequency predicted by Zipf for $k = 100,000$,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

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.

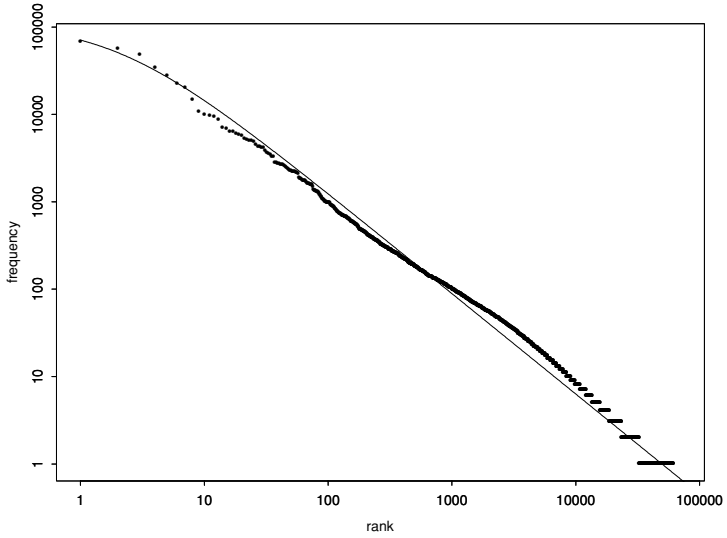


Figure 1.2 Mandelbrot's formula. The graph shows rank on the X-axis versus frequency on the Y-axis, using logarithmic scales. The points correspond to the ranks and frequencies of the words in one corpus (the Brown corpus). The line is the relationship between rank and frequency predicted by Mandelbrot's 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:

$$m \propto \sqrt{f} \tag{1.17}$$

or, given the previous law, that:

$$m \propto \frac{1}{\sqrt{r}} \tag{1.18}$$

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, 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

KEY WORD IN CONTEXT

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

Table 1.5 Frequent bigrams after filtering. The most frequent bigrams in a 1990 *New York Times* corpus 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
 7  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*.

$$\begin{array}{l}
\text{NP}_{agent} \text{ showed off (PP[}\textit{with/in}\text{]}_{manner}\text{)} \\
\text{NP}_{agent} \text{ showed (NP}_{recipient}\text{)} \left(\left\{ \begin{array}{l} \text{NP}_{content} \\ \text{CP[}\textit{that}\text{]}_{content} \\ \text{VP[inf]}_{content} \\ \textit{how}\ \text{VP[inf]}_{content} \\ \text{CP[}\textit{where}\text{]}_{content} \end{array} \right\} \right) \\
\text{NP}_{agent} \text{ showed NP[}\textit{interest}\text{] PP[}\textit{in}\text{]}_{content} \\
\text{NP}_{agent} \text{ showed NP[}\textit{aversion}\text{] PP[}\textit{to}\text{]}_{content}
\end{array}$$

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 *Tom 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

GRAMMATICALITY

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).

Lauer (1995b: 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, 1991), but many cognitive linguists would not endorse probability theory as a formalization of cognitive linguistics. See also (Schütze 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 4]

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.